Trabajo Práctico Final de Reglas de Asociacion

Keuroghlanian, Alejo
Krauthamer, Diego
Moncarz, Gabriel

***Resumen*—TO BE DONE**

***Index Terms*—Minería de datos Reglas de asociación Peliculas**

I. INTRODUCUCCIÓN

TO BE DONE

II. DATOS

En la presente sección se analiza losdatos originales, como fueron enriquesidos, como se los ha manipulado y las transformaciones realizadas. También se explica como todo el procesamiento se ha automatizado con intenciones de que este mismo análisis pueda realizarse en caso de tener tanto datasets distintos, como se desee incorporar nuevas entradas a los datos existentes.

II-A. Fuente de datos

El punto de partida del presente trabajo son archivos relacionados con la información sobre pelícuas y rating de películas. Los archivos en cuestión son 3

- Peliculas: contiene 3.883 películas a ser analizadas. La información sobre cada una de ellas es:
  - Código de película
  - Nombre de la película
  - Lista de géneros a la que pertenece

```
1::Toy Story (1995)::Animation
                |Children's|Comedy
2::Jumanji (1995)::Adventure
                |Children's|Fantasy
3::Grumpier Old Men (1995)
                ::Comedy|Romance
4::Waiting to Exhale (1995)
                ::Comedy|Drama
5::Father of the Bride Part II
                (1995)::Comedy
```

- Usuarios: contiene 6.040 registros de usuarios distintos. La información de cada fila es
  - código de usuario
  - sexo (masculino o femenino)
  - rango de edad
  - ocupación o profesion
  - código postal del usuario[1]

[1]Todos los usuarios analizados residen en los Estados Unidos

```
1::F::1::10::48067
2::M::56::16::70072
3::M::25::15::55117
4::M::45::7::02460
5::M::25::20::55455
```

- Rating: raring de distintos usuarios sobre diversas películas.[2]
  - Usuario que realiza la evaluación
  - Pelicula evaluada
  - Rating/evaluación asignada
  - tiempo: momento en que es evaluada

```
1::1193::5::978300760
1::661::3::978302109
2::3108::3::978299712
2::3035::4::978298625
5::3081::3::978243054
5::377::4::978245999
```

Estos 3 archivos son, en resumen, los datos crudos utilizados. Todo el tratamiento de datos posterior, enriquesimiento de información como los análisis y conclusiones finales, estan basados en el contenido de estos 3 archivos fuentes.

II-B. Preprocesamiento

TO BE DONE

II-C. Discretización de datos

TO BE DONE

II-D. Obtención de datos externos

TO BE DONE TODO: hablar del cache de peliculas

II-E. Proceso de automatización

TO BE DONE

II-F. Generación de datasets

TO BE DONE TODO: Hablar del formato de los datasets (Por transacciones). TODO: Explicar todos los datasets que se generaron TODO: Explicar que los datasets se generan dinámicamente

[2]Cabe destacar que cada usuario da su calificación sobre las películas que él desea. El usuario no esta obligado a completar una cantidad mínima de calificaciones.

II-G. Resumen

TO BE DONE

III. Procesamiento de datos

En esta sección se describe como se procesan los datos generados en la sección previa, con expectativas de encontrar reglas de asociación que permita hacer el analisis exploratorio objetivo.

En primera instancia, con intenciones de tener una idea general del problema, se intenta cargar los set de datos más genericos generados. Se busca explorar que reglas genera, que tipos de reglas, la cantidad de reglas generadas, como también cuales son los valores de soporte y confianza que retorne una cantidad de reglas aceptables y útiles.

III-A. Weka 3.7.11

Se intenta hacer el analisis de reglas de asociaciones con Weka. Este sofware por defecto no lee archivos de transacciones, donde una transacción puede ser representada en varias filas, sino que usa el formato donde todas las variables esta definidas en una fila. Para convertir de un formato a otro, se utiliza el filtro Denormalize[3] [4].

Una vez que se tienen los datasets listo para ser procesados en Weka, se intentan generar reglas de asociación exploratorias con Apriori. Se comienza con niveles de soporte y confianza elevados, aproximadamente de 0,9 en ambos, y como no generan reglas, se comienza a reducir ambos parámetros en 0,10. A medida que el algorítmo comienza a generar reglas, los tiempos de ejecución aumentan de forma exponencial. Esto ha impedido generar reglas de asociación no-triviales. Para obtener 5 reglas básicas, Weka demanda más de 1 hora de procesamiento. Esto complica realizar un análisis exploratorio supervisado por el usuario, objeto de interes en esta desarrollo.

Intentando descubrir los motivos de la lentidud de Weka, se han observado 2 factores fundamentales:

- Weka utiliza sólo 1 de los 4 nucleos disponibles, sin paralelizar el procesamiento
- El filtro Denormalize genera una matriz esparsa con Verdaderos y Falsos, pero **no** con datos faltantes para los falsos.

El primer punto no tiene salvación simple. Agregar multiprocesamiento al módulo de apriori esta fuera del alcanze del presente análisis.

El segundo punto tiene dos inconvenientes. Por un lado, como ya fue explicado, es el tiempo de procesamiento que demanda este formato. El segundo es que la mayor parte de las reglas de asociación generadas, son relacionadas a la no existancia de items, asunto que tiene sentido común.

Para resolver este segundo inconveniente se aplica un segundo filtro donde aplica datos faltantes a los registros falsos. De esta forma se tiene una representación precisa del problema. Al intentar hacer esto, Weka en primer instancia se queda sin memoria. Para ello se modifican los parámetros de la virtual machine de Java para asignarle mas memoria. Con 1 GB de memoria, el procesamiento de Apriori devuelve reglas de asociación básicas en un tiempo razonable (cercano a los 30 segundos). Sin embargo, cuando a Weka se le piden más reglas de asociaciones y se le disminuyen los rangos de soporte y confianza, el algoritmo requiere una cantidad de tiempo superior a una hora. Estas unidades de tiempo de procesamiento dificultan poder hacer un análisis introductorio al problema general, motivo por el que se decide intentar con otro software.

[3] http://weka.sourceforge.net/packageMetaData/denormalize/index.html

[4] El filtro Denormalize requiere la versión de Weka 3.7.2. Este es el motivo por el que el procesamiento se usa mediante una versión en desarrollo de Weka y no una estable

III-B. R 3.1.0

A la problemática inicial se le suma obtener resultados en tiempo razonables.

Se procede a realizar el mismo procesamiento de Weka usando R. Se usa el módulo arules. Éste módulo soporta que las entradas tengan un formato de transacciones[5] como de basket[6]. Durante todo el procesamiento y análisis se usa el formato por transacciones.

R muestra ser eficiente a la hora de generar reglas de asociaciones. R demora 130 segundos para cargar inicialmente la matriz esparsa de datos. Una vez que el dataset esta cargado, todas las corridas de Apriori se han ejecutado en menos de 1 segundo. El dataset de referencia es el dataset mas grande utilizado durante el trabajo de investigación. Éste tiene

- 8.870.000 líneas
- 888.000 transacciones
- peso: 220 MB
- contenido: información de películas, usuarios, actores y directores.

En primer termino genera desconfianza que los tiempos de ejecución de Apriori en R sean tan rápidos, más aún al contrastarlos con Weka. El motivo es porque arules maneja eficientemente la matriz esparsa de transacciones y porque no almacena los items inexistentes como lo hace Weka.

III-C. Elección de los parámetros de Apriori

Tener salidas de apriores con tiempos tan cortos, permite que puedan analizarse varios dataset, con distintos contenidos, y poder variar los parámetros de apriori gradualmente hasta conseguir los valores adecuados que generen reglas de asociación interesantes.

Esto permite hacer un análisis exploratorio inicial sin mayores dificultades. Más aún, todo el análisis del presente trabajo puede hacerse corriendo Apriori sobre todos los datasets generados con un bajo soporte, una baja confianza, y luego ordenar la salida por soporte y confianza. Una vez que se tiene las reglas de salida, sólamente resta seleccionar las mejores reglas no triviales con mejores soporte y confianza conjuntos.

Pese a que la librería arules tiene un exelente desempeño en sus tiempos de ejecución, se intenta encontrar los parámetros de Apriori adecuados, para que cuando se tenga un set de datos

[5] formato donde un campo define el número de transacción y otro campo define el id del producto

[6] formato donde en una fila se listan todos los productos de la transacción

tan pesado que el algoritmo demore un tiempo considerable en ser procesado, solo sea necesario hacer unas pocas iteraciones y no iterar masivamente hasta encontrar reglas de asociaciones interesantes. En pocas palabras, se intenta hacer un análisis exploratorio inteligente y no de fuerza bruta.

Para esto se utiliza el paquete arulesviz, librería que ayuda a mostrar las distintas reglas en un gráfico de dispersión. El gráfico muestra las reglas en los distintos niveles de soportes y confianzas. Entonces, un análisis inteligente puede hacerse corriendo una vez Apriori con bajos niveles de soporte y confianza, y en base a la nube de puntos se puede elegir estos argumentos para que genere una cantidad aproximada de reglas deseada. Luego, se deben observar las reglas mismas para ver si los resultados son los esperados, y en caso de ser necesario, correr nuevamente Apriori con valores mas ajustados u holgados.

En el grafico 1 muestra la nube de puntos de uno de los dataset de películas. Al comenzar a estudiar reglas de asociación se tiene la expectativa de poder generar reglas con niveles de soporte cercanos al 90 %, pero tan solo conver el gráfico se puede apreciar que hay muy pocas reglas con soporte mayor al 10 %, y que la mayoría estan recién cercanas al 0,5 %

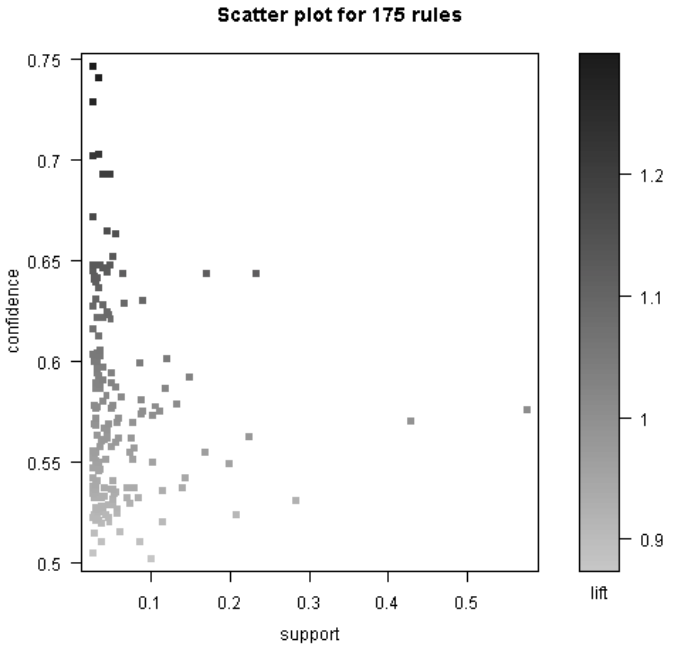


Figura 1. Gráfico de dispersion de reglas generadas

Al correr Apriori con un minsup del 10 %, para ver cuales son estas reglas tan fuertes respecto de las demas, se observan que son reglas cuyo lado izquiero esta vacío, o reglas tan triviales como {sexo=hombre} => {rating=high}

Es por esto que, para el análisis de peliculas realizado en este informe, se han realizado, en primera instancia, corridas de Apriori con soportes mínimos en el rango de

$$0,009; 0,03$$

y confianza en el rango de

$$0,05; 0,1$$

IV. Análisis

En esta sección se analizan las mejores reglas generadas en la sección de procesamiento de datos. Se muestran reglas que representan gustos de películas, se intenta establecer reglas relacionadas con la posición demográfica, como también analizar similitudes entre los los rankings de peliculas en el año 2000 y 2001.

IV-A. Reglas generales

La primera impresión al ver las reglas generadas por Apriori, es que todas referencian a ratings positivos, siendo muy pocas las relacionadas con calificaciones mediocres o bajas. Es por esto que se analiza la frecuencia de estas 3 calificaciones en la tabla I, donde puede claramente notarse que existe una tendencia a calificar las películas de forma elevada.

Cuadro I
Frecuencia de ratings

Calificación	Frecuencia
low	16 %
medium	26 %
high	58 %

IV-A1. Directores mejor rankeados: La tabla II muesta las reglas relativas a los directores de películas mejor rankeados.

Cuadro II
Reglas relativas a directores

regla	sop.	conf.	lift
{Steven Spielberg} => {high}	2,36 %	90 %	1,15
{Alfred Hitchcock} => {high}	1,23 %	95 %	1,22
{James Cameron} => {high}	1,20 %	89 %	1,14
{Rob Reiner} => {high}	1,17 %	92 %	1,18

El soporte del director mejor rankeado, Steven Spielberg, es de un 2,36 %. Aunque a nivel porcentual este soporte parece bajo, a nivel nominal significa que de 888.500 ratings totales, en 21.000 aparecen Steven Spielberg y de esas 18.900 la calificaron como un rating {high}. Los soportes de la tabla II son nominalmente significativo.

IV-A2. Actores mejor rankeados: La tabla III muestra las reglas relativas a los actores mejor ranqueados.

Cuadro III
Reglas relativas a actores individuales

regla	sop.	conf.	lift
{Harrison Ford} => {high}	2,85 %	91 %	1,17
{Tom Hanks} => {high}	1,96 %	89 %	1,15
{Robert De Niro} => {high}	1,53 %	87 %	1,11
{Sean Connery} => {high}	1,35 %	85 %	1,09
{Arnold Schwarzenegger} => {high}	1,17 %	77 %	0,99

IV-A3. dupla de actores mejor ranqueados: La tabla IV muestra las reglas relativas a la dupla de actores mejor ranqueados.

Esta tabla confirma aún mas que Harrison Ford es uno de los actores preferidos.

Cuadro IV
REGLAS RELATIVAS A DUPLA DE ACTORES

regla	sop.	conf.	lift
{Carrie Fisher, Harrison Ford} => {high}	1,10 %	95 %	1,22
{Harrison Ford, Mark Hamill} => {high}	1,10 %	95 %	1,22
{Billy Dee Williams, Carrie Fisher} => {high}	0,70 %	94 %	1,20

Cuadro V
REGLAS RELATIVAS A ACTORES INDIVIDUALES

regla	sop.	conf.	lift
{Drama} => {high}	23 %	64 %	1,11
{Comedy} => {high}	20 %	54 %	0,95
{Action} => {high}	14 %	53 %	0,93
{Romance} => {high}	9 %	57 %	0,99

IV-A4. Generos: La tabla V muestra los generos mejores ranqueados.

La base de datos contiene 18 categorias. En caso de que haya independencia en los gustos de géneros, el soporte de cada categoría debería tender a 5,55 %. Estos niveles de soporte, superiores al 10 % y algunos cercanos al 20 %, implican que hay una preferencia sobre los géneros de drama, comedia, accion y romance.

IV-A5. A los hombres les gusta las peliculas de guerra: La tabla VI muestra que a los hombres les gusta las películas de guerra, con o sin drama. Sin embargo, hay una sensible preferencia por las peliculas de guerra sin drama.

Cuadro VI
REGLAS RELACIONADAS CON PELICULAS DE GUERRA

regla	sop.	conf.	lift
{male, war} => {high}	3,75 %	69 %	1,20
{Drama, male, war} => {high}	2,51 %	75 %	1,29

IV-A6. El declive de la década de los ' 90: Se han generado pocas reglas con calificaciones negativas. Una de las que llama la atención es que muchas de ellas son relacionadas con películas de la década de los ' 90, tal como muestra en la tabla VII.

Cuadro VII
EL DECLIVE DE LOS ' 90

regla	sop.	conf.	lift
{25−34, low, male} => {1990}	3,44 %	64 %	1,20
{Comedy, low, male} => {1990}	3,07 %	65 %	1,21
{Drama, low} => {1990}	2,77 %	67 %	1,21
{Action, low, male} => {1990}	2,74 %	66 %	1,25

IV-A7. Las personas de 45 a 55, son quienes mejores califican: La tabla VIII muestra las reglas de las edades de los que mejor califican. Probablemente no sea casualidad que ambas categorías son consegutivas.

Cuadro VIII
MEJORES CALIFICACIONES SEGÚN LA EDAD

regla	sop.	conf.	lift
{45−49} => {high}	4,50 %	62 %	1,08
{50−55} => {high}	4,49 %	59 %	1,03

IV-A8. Generos en común: La tabla IX muestra generos relacionados entre si.

Cuadro IX
GENEROS EN COMÚN

regla	sop.	conf.	lift
{Action,high,Sci-Fi} => {Adventure}	2,53 %	51 %	3,80
{high,Sci-Fi,Thriller} => {Action}	1,93 %	76 %	2,94

IV-B. Reglas demográficas

En esta sección se analiza las relaciones y preferencias según las variables demográficas estado y ciudad.

IV-B1. Estados y peliculas: La tabla X hace una comparación entre que porcentaje de votos hicieron los estados que más calificaron y el porcentaje de habitantes que tiene el estado respecto del total del país.

Cuadro X
VOTOS Y PORCENTAJE POBLACIONAL

estado	% votos	%población
California	18,07	11,78
New York	7,08	6,13
Minnesota	6,45	1,70
Illinois	5,50	4,07
Texas	5,23	8,36

Estos resultados muestran, principalmente, que tanto el estado de California como el de Minesotta, son outlayers severos. Mientras que la población de Minesota es del 1,70 % de la población de USA, el 6,45 % de los votos pertenecen a este estado. El ratio es cercano a 4 veces más su población, lo que demuestra que el estado de Minnesota tiene una preferencia de entretenerse viendo películas. Algo similar sucede con el estado de Californa, donde su población es de casi el 12 % de los Estados Unidos, pero el 18 % de los votos totales pertenecen a este estado. California también es un outlayer, pero este se comprende mas ya que la mayor industria cinematográfica de Estados Unidos es Hollywood, que se encuentra en el estado de California.

Otro de los puntos destacados que muestra la table X es que el estado de Texas no es propenso a entretenerse mirando películas.

Otra regla interesante es que el estado que mejor califica películas es California.

regla	sop.	conf.	lift
{high} => {CA}	10,44 %	57 %	1,03
{MN} => {HIGH}	3,81 %	59 %	1,03

Este sesgo posiblemente también se deba a que Hollywood pertenece a California, y mucha gente que vive de esta industria o se ve influenciada por su cercanía física, esta calificando de forma positiva las películas. Es consistente que la regla {MN} => {HIGH} tiene un soporte del 3,81 %, respecto del 10,44 % de California, ya que Minesotta no se encuentra influenciada por Hollywood.

IV-B2. ciudades y peliculas: La tabla XI muestra las ciudades que más votos realizaron y la tabla XII muestra las reglas de ciudades con mas calificaciones positivas. Ambas tablas son consistentes entre si, y también son consistentes con el análisis de los estados. Lo que llama la atención es que la ciudad de Los Angeles, la 2da ciudad mas pobladas de los

Estados Unidos, ciudad que también pertenece al estado de California, **no** esta incluida entre las que mas votan[7], como tampoco se pudieron generar reglas que incluyan a esta ciudad y a algúna calificación. Esta inexistencia de reglas para esta ciudad particular, implicaría una "pseudo-regla"que refleja que a la ciudad de Los Angeles **no** le gusta ver películas.

Cuadro XI
VOTOS POR CIUDADES

ciudad	% votos %
Minneapolis	2,52 %
New York	2,37 %
San Francisco	2,12 %
Saint Paul	2,09 %

Cuadro XII
VOTOS POSITIVOS POR CIUDAD

regla	sop.	conf.	lift
{Minneapolis} => {high}	1,48 %	58 %	1,02
{New York} => {high}	1,38 %	58 %	1,01
{San Francisco} => {high}	1,22 %	57 %	0,99

IV-B3. Estados y géneros: En esta sección se analiza la relaciones existentes entre estados y géneros de películas. La tabla XIII resume la reglas mas relevantes al respecto.

Cuadro XIII
REGLAS MAS RELEVANTES SOBRE GÉNEROS Y ESTADOS

nro	regla	sop.	conf.	lift
I	{} => {Drama}	36 %	36 %	1
II	{} => {Comedy}	35 %	36 %	1
III	{} => {Action}	26 %	26 %	1
IV	{} => {Thriller}	19 %	19 %	1
V	{CA} => {Drama}	6,69 %	37 %	1,02
VI	{CA} => {Comedy}	6,32 %	34 %	0,97
VII	{CA} => {Action}	4,70 %	26 %	1,00
VIII	{CA} => {Thriller}	3,48 %	19 %	1,14
IX	{CA,high} => {Drama}	4,31 %	41 %	1,42
X	{CA,high} => {Comedy}	3,53 %	33 %	0,94

La 1ra a 4ta regla muestra que en todo Estados Unidos, la gente mira príncipalmente pelícuas de drama y comedia, en 3er medida de acción y en 4ta Thriller. La 4ta a 8va regla se muestran las asociaciones mas fuertes que del lado izquierdo pertenece a un estado, y el derecho a un género de película. Las 4 reglas tienen un soporte relativamente significativo y pertenecen todas al mismo estado: California, estado con mayores votaciones. Lo más interesante, es que los géneros de películas más vistos en California, coinciden con los de todo Estados Unidos, tanto en orden, como en distancia relativa entre cada uno de ellos. La 9ena y 10ma regla muestra las asociaciones con mayores soporte y confianza, cuyo lado izquierdo tiene un estado y una calificación elevada, y el lado derecho un género. Los resultados, no por casualidad, coinciden con las películas mas vistas en California, como en todo Estados Unidos.

El análisis de la tabla XIII nos permite concluir que California es un estado dominante en cuestión de género de peliculas.

Sería interesante analizar si esta misma posición dominante se repite en el pasado. Si llegase a ser así, pero con diferentes géneros, se podría pensar que la industria cinematográfica de Holliwood sigue los gustos de California, y un cambio en los gustos de ese estado, impactaría en el tipo de peliculas que Estados Unidos produciría.

La tabla XIV muestra las reglas de asociación mas fuerte que destacan los géneros mejores calificados para cada estado relevante.

Cuadro XIV
GENEROS MEJORES CALIFICADOS POR ESTADO

regla	sop.	conf.	lift
{CA,Drama} => {high}	4,31 %	64 %	1,11
{CA,Comedy} => {high}	3,53 %	56 %	0,97
{Drama,NY} => {high}	1,79 %	63 %	1,13
{Comedy,NY} => {high}	1,36 %	54 %	0,98
{Drama,MN} => {high}	1,50 %	65 %	1,07
{Comedy,MN} => {high}	1,37 %	56 %	0,88
{Drama,IL} => {high}	1,21 %	61 %	1,13
{Comedy,IL} => {high}	1,01 %	50 %	0,98
{Drama,TX} => {high}	1,22 %	65 %	1,15
{Comedy,TX} => {high}	1,11 %	56 %	0,98
{Drama,MA} => {high}	1,12 %	66 %	1,15
{Comedy,MA} => {high}	1,00 %	58 %	1,02

Esta tabla refuerza aún más la posición dominante que tiene California sobre la industria cinematográfica. En todos los estados mas destacados, los géneros de las peliculas mas vistas coinciden con los de California. La diferencia es el nivel de soporte de estas reglas respecto a porcentaje poblacional e inclusive al porcentaje de votos en esta encuenta. California tiene porcentajes relativos mucho mas fuertes, avalando la hipótesis de que Hollywood produce películas del agrado de California, y el resto de los estados mira las películas realizadas por su mayor productor.

IV-B4. Ciudades y géneros: Al intentar encontrar la existencia de relaciones entre ciudades y géneros, se han encontrado las reglas de la tabla XV.

Cuadro XV
GENEROS MAS VISTOS POR CIUDAD

regla	sop.	conf.	lift
{New York} => {Drama}	0,99 %	42 %	1,63
{Minneapolis} => {Drama}	0,93 %	37 %	1,02
{Seattle} => {High}	1,12 %	63 %	1,10

Estas reglas no aportan información relevante, solo refuerzan el desarrollo previo. Posiblemente la imposibilidad de encontrar una fuerte relación entre ciudades y géneros se debe a 2 motivos:

- El nivel de granularidad de agrupamiento por ciudad es elevado.
- Los géneros de películas mas vistas estan muy concentrados. Los 3 géneros mas importantes se llevan más del 70 % de la torta, haciendo insignificante la producción de otros géneros. Si los géneros producidos fuesen estuviesen distribuidos más uniformemente, posiblemente pueda esperarse que a la ciudad de Utah tenga preferencias por las películas religiosas, los de New York por musicales y Las Vegas por películas para adultos.

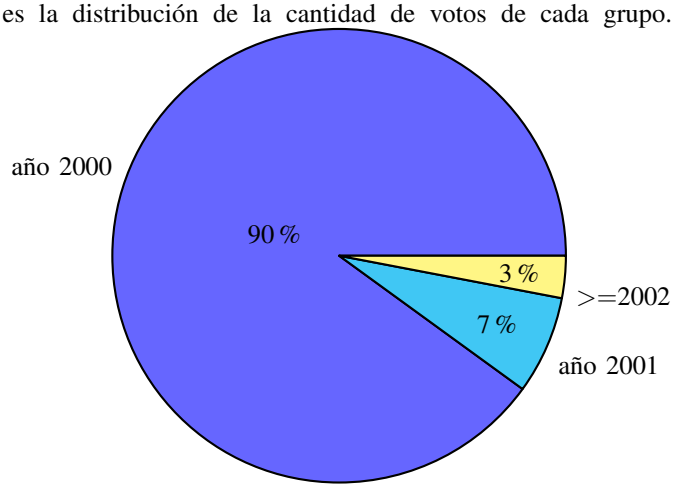
IV-B5. Análisis de usuarios: En esta sección se analiza las carácteristicas de los usuarios que realizan calificaciones. Con este propósito se divide el set de datos en 3 grupos:

[7]Los Angeles tiene un 1 % de los votos totales

1. Los usuarios que calificaron películas en el año 2000.
2. Los usuarios que calificaron películas en el año 2001.
3. Los usuarios que calificaron películas desde el 2002 para adelante.

El objetivo es encontrar características similares de los primeros 2 grupos y verificar si éstas también se encuentran presentes en el último.

Para este análisis lo que primero se tiene en cuenta es la distribución de la cantidad de votos de cada grupo.



El 2do paso a realizar es hacer algunas corridas exploratorias en cada grupo y graficar la nube de puntos con el soporte de arulesViz.

En las figuras 2, 3 y 4 se puede ver que, con los mismos parámetrosi, la cantidad de reglas generadas son del mismo orden, como también, la distribución de las reglas en la nube de puntos son muy similares en los 3 grupos.

Al navegar detenidamente la reglas más importantes de los 3 grupos se puede apreciar la semejanza entre ellas. La tabla XVI hace un resumen de estas.

Las primeras 2 reglas muestran el sexo de los usuarios. En los 3 conjuntos de datos los hombres han tenido una posición mayoritaria holgada. Sin embargo, a partir del 2002, las mujeres tienen una mayor participación en el set de votaciones, lo que se puede inferir 3 hipótesis no excluyentes:

- La mujer esta teniendo mas particion en el rubro informático y/o acceso a Internet, ya que las votaciones se hacen on-line
- Las mujeres estan viendo mas películas a partir del año 2002.

Para obtener la 3er regla se solicita a R que devuelva la profesión que más votaciones ha realizado en cada set de datos, siendo los administrativos los primeros en todos los casos. Dado que hay más de 20 profesiones, en caso de haber independencia de los gustos sobre profesiones, el valor esperado sería del 5 %, pero los administrativos tienen valore comprendidos entre el 13 y 17 %, lo que comprueba que los administrativos son los más afines a entretenerse mirando películas.

Luego se hace un análisis por edad. Al analizar este area, se debe tener conciencia sobre el entorno y sobre la posibilidad de acceso de las distintas edades al sistema de calificaciones. Es posible que los chicos pequeño vean muchas películas, pero no tengan capacidad de realizar votaciones. Del mismo modo, es posible que las personas muy adultas y retiradas también sean asiduos expectadores, pero no tengan interes en contribuir al sistema de rating. Por lo tanto, hay que tomar conciencia que los análisis por edades pueden estar sesgado por el interes de una generación en brindar información.

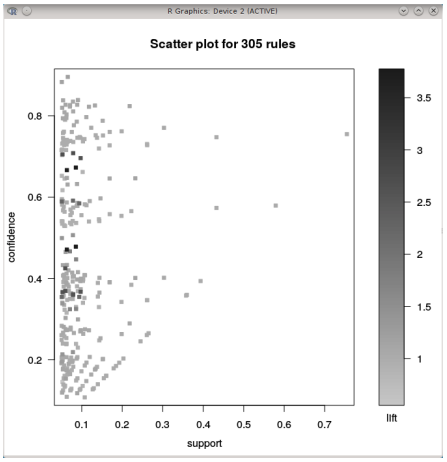


Figura 2. Reglas para el año 2000. Minsup=0.05 Minconf=0.1

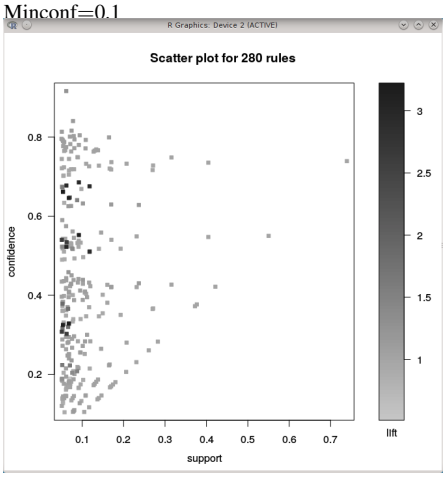


Figura 3. Reglas para el año 2001. Minsup=0.05 Minconf=0.1

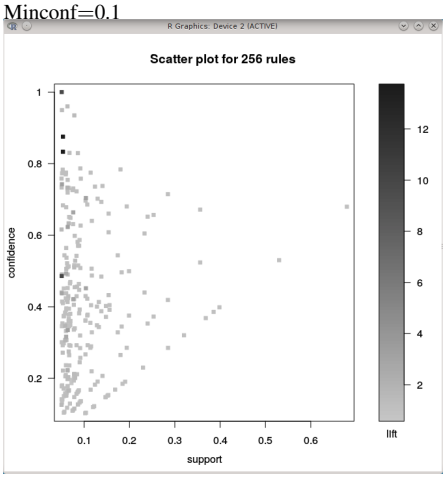


Figura 4. Reglas para el año 2002 en adelante. Minsup=0.05 Minconf=0.1

Cuadro XVI
REGLAS DE SEMEJANZA DE USUARIOS

nro	regla	año 2000		año 2001		año 2002	
		sop	conf	sop	conf	sop	conf
I	{} => {male}	75 %	75 %	74 %	74 %	67 %	67 %
II	{} => {female}	24 %	24 %	26 %	26 %	32 %	32 %
III	{} => {Administ.}	13 %	13 %	17 %	17 %	15 %	15 %
IV	{} => {25−34}	39 %	39 %	42 %	42 %	40 %	40 %
V	{} => {35−44}	20 %	20 %	17 %	17 %	14 %	14 %
VI	{} => {18−24}	18 %	18 %	23 %	23 %	22 %	22 %
VII	{} => {high}	58 %	58 %	55 %	55 %	53 %	53 %
VIII	{} => {medium}	26 %	26 %	28 %	28 %	28 %	28 %
IX	{} => {low}	16 %	16 %	16 %	16 %	18 %	18 %
X	{H. Ford} => {high}	2,18 %	74 %	1,33 %	66 %	1,25 %	65 %
XI	{18−24} => {Administ.}	8,54 %	47 %	11 %	50 %	10 %	70 %
XII	{Advent} => {Action}	9,72 %	69 %	6,77 %	64 %	6,37 %	62 %
XIII	{romance} => {25−34}	6,05 %	39 %	6,09 %	41 %	6,06 %	39 %
XIV	{Administ.} => {high}	7,23 %	56 %	9,10 %	52 %	6,86 %	46 %
XV	{S. Spielberg} => {high}	1,81 %	73 %	1,10 %	72 %	1,07 %	68 %

La regla IV muestra que el rango de 25 a 34 años es el que, por lejos, más peliculas mira. En todo momento lleva una diferencia de aproximadamente el doble respecto al segundo.

Las reglas V y VI muestran que hay un cambio de tendencia con los rangos 35−44 y 18−24: A partir del 2001, el rango 18−24 tiene una participación más activa al calificar películas que el de 35−44, ya que pasan de un 18 % en 2000, a un 23 % y 22 % a partir del 2001.

Las reglas VII, VIII y IX muestran que tan bondadosos son los usuarios. En todo momento los usuarios tienden a calificar positivamente las películas y muestran rechazo por asignar calificaciones bajas. Las calificaciones positivas son constantemente más del doble de las mediocres y más del triple las negativas. Sin embargo, con el tiempo se va observando que el espectador se vuelve un poco más exigente, ya que en el año 2000 los votos positivos eran un 58 %, en 2001 un 55 % y desde el 2002 un 53 %.

Para generar la regla X se solicita el actor mejor calificado en cada dataset. En todos estos Harrison Ford esta liderando el ranking.

Para generar la regla XI, se pide a R que retorne las profesiones más relevantes para cada rango de edad, siendo {18−24} => {Administrator} las de mayor soporte en todo momento.

La regla XII muestra que las personas que miran películas de aventuras, también observan de acción. Esta regla se conserva en los 3 períodos analizados.

La regla XIII se analiza cual es el rango de edad preferido para las películas románticas. En todos los set de datos fue el rango de 25−34, con un nivel de soporte y confianza parejo en todo momento. Esto podría inducir que, por cuestiones psicológicas, los jovenes de 25 a 34 años son los más románticos.

Para generar la regla XIV Se le pide al software que devuelva cual es la profesión más generoso al calificar películas. Los administrativos son lideres en este aspecto con el correr del tiempo.

Para la regla XV se pide el director de cine mejor calificado en cada período. Steven Spielberg estuvo en la vanguardia en todo momento.

Por último, en la table XVII se analiza cual es el género mas visto por las mujeres en cada momento.

Cuadro XVII
GENEROS MAS ESCOGIDOS POR LAS MUJERES

year	regla	sop.	conf.
2000	{Romance} => {female}	5,22 %	33 %
2001	{Comedy} => {female}	10,12 %	27 %
>= 2002	{Romance} => {female}	6,33 %	41 %

Tanto en el año 2000 como a partir del 2002, es el género romántico. Sin embargo, en el año 2001, el género de comedia prima por sobre el romance en más de 100 %: en 2001 y desde 2002, el soporte de {Romance} => {female} es de entre un 5 y 6 %, pero en 2001, la regla {Comedy} => {female} es de un 10 %.

Previo a finalizar esta sección, debe recordarse y tenerse en cuenta que el 90 % de los datos que están en el dataset son del año 2000, quedando solo un 6 % para el año 2001 y un 3 % para los años posteriores. Pese a que un 3 % no es un porcentaje significante, son 24.000 votaciones, número no despreciable. Como se analizaron cada rango de tiempo en dataset independientes, el soporte y la confianza de las reglas son comparables entre si, independientemente de la cantidad nominale de datos, ya que estos resultados son porcentuales.

V. CONCLUSIONES

TO BE DONE

VI. POSIBLES MEJORAS

Normalizar las calificaciones POR USUARIO Hablar de crear una DB Hablar de probar con Neo4J

REFERENCES