Visualización de Información CC5208
Departamento de Ciencias de la Computación Facultad de Ciencias Físicas y Matemáticas Universidad de Chile

Tarea 3

Alumnos: Martín Panza
Juan Paulsen

Profesor: Javier Bustos

Ayudante: Diego Madariaga

Fecha: 09-11-2016

0. Dataset a Utilizar

Habiendo tomado en cuenta la alta cantidad de datos faltantes que poseía el dataset de Ranking de Universidades utilizado para la tarea 1, junto con las limitaciones que se tenían para relacionar los datos de cada ranking (Sólo coincidían en 4 años por ejemplo); se decidió cambiar el dataset a utilizar, puesto que los problemas anteriores limitaron ciertas visualizaciones e impedían realizar muchas otras. Así, a partir de esta tarea se trabajará con un dataset acerca de películas puntuadas por usuarios en la web.

1. Descripción de los Datos

Se recopilaron dos sets de datos de fuentes diferentes: Kaggle, sitio que recopila set de datos para investigación; y MovieLens, sitio web sin fines comerciales de recomendaciones de películas. Ambos contienen datos sobre variadas películas desde 1916 en el primer caso, e inclusive cortometrajes que datan del siglo XIX en el segundo caso; hasta películas de la actualidad (2016) en ambos datasets. En particular, ambas poseen una puntuación asignada por los usuarios de la página correspondiente. En el caso de Kaggle, de la página web de IMDB, mientras que MovieLens de su propia página. Además poseen otros atributos particulares:

-El primer dataset de Kaggle cuenta con 5043 registros de películas almacenados en un archivo con 28 atributos entre los que se encuentran: nombres de actores y de director, idioma, país, géneros, link en IMDB y título como datos de categoría exclusiva; rating de contenido como categoría con orden; año, que se considera ordinal; y duración, presupuesto, recaudación, puntuación en IMDB y likes en facebook de actores, director y película como datos ordinales con cero absoluto.

-El otro set de datos perteneciente a MovieLens contiene 34208 registros de películas. Cada película tiene asociada un id propio del dataset junto a su id en IMDB e id en *the movie database*; ambos categoría con orden. Además, cada película puede tener asociados varios géneros, tags y ratings. Tanto tags como ratings son registrados por usuarios, por lo que a su vez contienen el id del usuario y el tiempo en que se hizo el registro. El número de registros es 586994 para tags y 22884377 para ratings. Géneros y tags son categoría exclusiva. Los id's de usuario y películas son datos de categoría con orden. Por último, ratings y tiempo son ordinales y con cero absoluto.

El dataset de Kaggle posee atributos con algunos datos faltantes (NA). Sin embargo, para cada atributo, este número no sobrepasa los 50 registros. A excepción de los atributos que tienen que ver con temas monetarios; es decir, presupuesto y recaudación, que alcanzan 800 NA's. A su vez, MovieLens solo posee datos faltantes en el id de *the movie database* (296).

Con el fin de obtener el máximo de atributos para cada película, se cruzaron ambos dataset de acuerdo al link en imdb de Kaggle, que posee a su vez el id de imdb que se encuentra también en MovieLens. Este join resultó en un dataset de 4566 con los 28 atributos de Kaggle y agregando los tags y géneros de MovieLens. Para ser consistente, se tomará en cuenta sólo la puntuación en imdb que se posee de Kaggle. Además de ciertos atributos que no serán considerados en la visualización por ser poco relevantes o ser imprecisos. Por último, algunos datos repetidos se filtraron, resultando finalmente en 4452 registros.

2. Preguntas Que Se Espera Responder

A) ¿Qué películas valen o no valen la pena ver, tomando en cuenta la relación duración-puntuación?
B) ¿Cuáles son las buenas películas en cada género?
C) ¿Qué otras películas tiene cada director/actor y cuál es su puntuación?
D) ¿Quiénes son los actores y directores que tiene cada película?
E) Detalles de las películas seleccionadas como año, idioma, rating, tags.
F) ¿Cuáles son los actores, directores y géneros con mejores puntuaciones?

3. Mapeo de Datos

A continuación se detalla el diseño del layout, representado en la figura 2 de la siguiente parte (4).

El objetivo de la visualización es permitir al usuario encontrar las películas mejor puntuadas de acuerdo a los atributos que busque. Ya sea categoría, rating de contenido, año, actores, etc. Dando también importancia a la duración de la película, suponiendo que el tiempo que la persona esté dispuesta a utilizar en ver la película sea en ocasiones limitado. En este sentido, lo intuitivo es hacer un gráfico de duración v/s puntuación en un scatterplot. Pero, dada la cantidad de registros es imposible visualizarlas todas al mismo tiempo. Resultaría algo como esto:

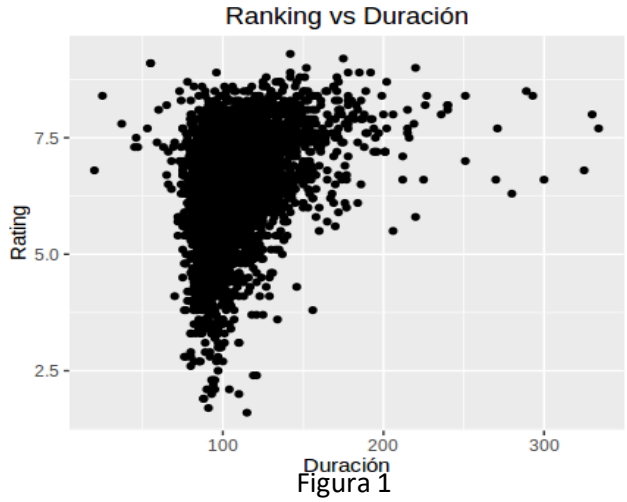

Figura 1

Así, es vital agregar un filtro que permita al usuario acotar la cantidad de películas mostradas en la visualización. Este filtro utilizará los atributos:
- Género
- Rating de contenido
- Año
- País

Para el filtro se pretende usar choicebox para elegir entre los atributos posibles. Excepto para año que es un dato ordinal, que se puede seleccionar en rango.

Además, ya que aún pueden mostrarse muchos registros en la visualización, sería muy conveniente permitir hacer zoom y navegar sobre la visualización. Esto para reducir o aumentar el rango de puntuación o el rango de duración que se visualiza.

De esta forma se respondería la pregunta B. Sin embargo, para responder C se utilizará un gráfico muy similar pero con diferente filtro, pues en el caso de los actores/directores nos podría interesar conocer con qué otros director/actores trabajó. Por lo que el filtro esta vez consistiría de choiceboxes con:

- Director	ó	-Actor1
- Actor1		-Actor2
- Actor2		-Actor3
(Caso en que se visualicen las películas de un actor)		(Caso director)

Para conectar a ambos gráficos, se permitirá la selección de una película en particular en ambos gráficos. Al seleccionar una película se mostrarán todos los detalles de esta, incluyendo título, géneros, año, rating, actores, puntuación, duración, color, etc… y en la que se podrá seleccionar nuevamente géneros, director, actores o director para mostrar inmediatamente el gráfico correspondiente. En adición, un nuevo gráfico de barras nos permitirá visualizar cuáles son los tags más importantes de la película seleccionada. Con esto se respondería las preguntas D y E.

Es probable que el usuario necesite marcar ciertas películas para no tener que buscarlas otra vez, por ejemplo si ya la vio, o si la marca para verla. Se debe permitir esta funcionalidad de anotación, y para eso se utilizarán colores. Se usará un color para marcas, además del color predeterminado que traerán los puntos del scatterplot y un último que nos muestre cual es la película que está siendo seleccionada. Los colores más convenientes serían verde, rojo y azul de acuerdo al modelo de color RGB.

En cuanto a la pregunta F, otra visualización aparte nos permitirá comparar los atributos indicados. Esta se compondrá de 4 diferentes gráficos. Dos de línea en que puedan identificarse los elementos mediante colores sobre los ejes Año vs Puntuación (escala no lineal) y Puntuación v/s Densidad. Junto a dos de barras que nos muestren mediana de puntuación, y cantidad de películas respectivamente.

Esta separación dentro de la visualización mejora nuestra expresividad ya que reducimos la información a mostrar en cada gráfico. Asimismo, el scatterplot nos muestra sólo el punto en que se encuentra la película en la relación duración-puntuación, lo que es positivo. Pero se puede tener acceso a todos los detalles mediante la selección. En este sentido, también existe una mejora debido al filtro, aunque esto depende más del usuario pues si muestra todos los puntos empeorará la expresividad. De la misma forma, los gráficos de comparación separan la información para mayor expresividad, pero es posible sobrecargar los gráficos con un mal uso del filtro.

A su vez, es el mismo scatterplot quien reduce nuestra eficacia en el sentido de que no podemos obtener información rápidamente para un registro en particular. No obstante, este problema se soluciona en gran medida con la funcionalidad de selección. También afecta el tener que cambiar de visualización para ciertos atributos como lo son los actores, aunque es necesario. Por otro lado, el gráfico de tags y los de comparación nos permiten obtener muy rápidamente la información que buscamos. Los detalles pueden mostrar relativamente numerosos atributos, pero no debiese impedir la eficacia.

Los gráficos son coherentes pues en el scatterplot ambos atributos son cuantitativos. Por su parte el gráfico de barras de tags muestra datos cualitativos versus datos cuantitativos. Esto mismo ocurre con las barras de los últimos dos gráficos en las comparaciones. Los gráficos de líneas por su parte utilizan ambos ejes con datos cuantitativos.

El uso de variados gráficos para mostrar mayor información es uno de los principios de Tufte que se cumple, y a excepción de las barras pintadas, se optimiza el uso de tinta de acuerdo a la razón Data/Ink. No se utilizan muchas variables ni imágenes como pide Tufte, pero sí se establecen comparaciones, se permite visualizar varios gráficos con información relacionada al mismo tiempo y se utilizan cantidades, de hecho el atributo más importante es la puntuación que representa un dato cuantitativo.

Por último, la separación nos permite reducir el número de focos de atención a un número menor que 7+/-2, lo que hace más fácil y no sobrecarga al usuario. Además se utilizan tamaños diferentes en los gráficos para captar más la atención del usuario. La idea es que se vea el scatterplot (de mayor tamaño), se encuentre un punto interesante, y cuando se quieran conocer sus detalles, se seleccione y se dirija la concentración a los detalles que se muestren en la lista que es más pequeña. También los tags funcionan como detalles. No se utilizan muchos colores para no perder la diferenciación. Aunque un mal uso del último filtro puede dar cabida a este problema.

4. Diseño del Layout

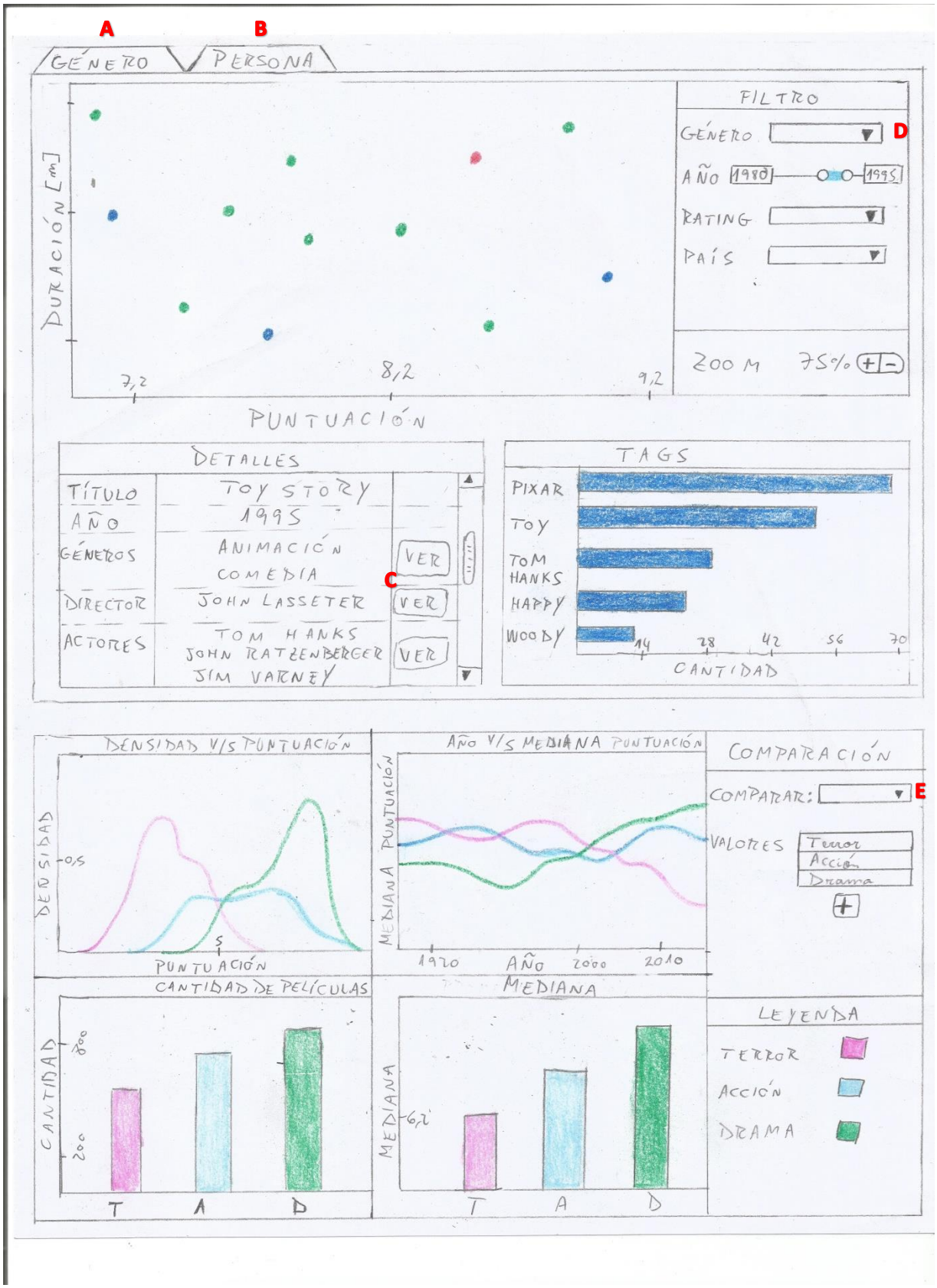

Figura 2

Como se puede observar en las marcas A y B de la figura 2, la primera parte de la visualización posee dos pestañas, en donde se tienen dos visualizaciones diferentes. La que se muestra en la imagen corresponde a la de géneros. Al cambiar a la de personas lo que cambia son los elementos del filtro. Aparte de A y B, la otra forma de cambiar de pestaña es dentro de los mismos datos de la película seleccionada (color rojo) como muestra la marca C. Aquí, al apretar el botón 'ver' que se encuentra en la línea de género, se nos llevará a la pestaña de género con la película marcada en la visualización filtrada por aquel género. De otra forma, los botones 'ver' que están en la línea de actores y directores, nos llevarán a la pestaña de personas con la película marcada en la visualización filtrada por tal actor o director y con un filtro con las siguientes choiceboxes:

-Actor 1/Director
-Actor 2/Actor 1
-Actor 3/Actor 2
(dependiendo del caso)

Además de un textbox con la persona actual.

Explicando un poco más los choiceboxes como se muestra la marca D, estos permiten el elegir de una lista finita de posibilidades, cero, una o más para añadir al filtro. Por ejemplo en género encontraremos: Terror, Acción, Comedia, Drama, Ciencia ficción, etc. O en el caso de los actores mostrará la lista actores o directores con los que ha coincidido en alguna película.

En cuanto a la marca E, ése choicebox tendrá tres opciones por las que se podrá comparar: género, actor y director. En 'valores', se podrá elegir cuáles registros se utilizarán para comparar, con la posibilidad de agregar más a la lista y actualizando la leyenda de forma correspondiente.

5. Repositorio en Github

https://github.com/martinpanza/Visualizacion-de-Informacion-Peliculas

6. Fuentes

https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset
https://movielens.org/