

# Sistema de Recomendación

Luis Murillo Vizcardo, Enrique Gutiérrez Salazar

7 de mayo de 2019

## 1. Almacenamiento y levantamiento de la base de datos

Para el almacenamiento de datos usamos una estructura del STL de C++ llamado UNORDERED MAP la cual a diferencia del MAP normal nos permite hacer consultas y y inserciones en un tiempo constante  $O(1)$  y en el peor de los casos seria  $O(n)$  en caso de colisión de claves, esto se debe a que almacena sus datos en una tabla hash mientras que el MAP normal almacena sus valores en un árbol balanceado de búsqueda binaria (BST) y su tiempo seria  $O(\log(n))$ . Por otra parte el UNORDERED MAP usa más memoria que un MAP normal.

## 2. Implementación del K-NN

Para la implementación de K-NN, hicimos el cálculo al momento de la lectura del archivo, calculando solo las variables requeridas para calcular las respectivas distancias. Como la información esta ordenada por el ID del usuario entonces la puntuación de cada usuario estará en un solo bloque lo cual es fácil de sacar la información, verificando que el ID del usuario no cambie, y si lo hace significa que comienza la puntuación de otro usuario. Las variables calculadas para cada distancia son las siguientes:

### 2.1. VARIABLES PEARSON

```
if(idU==idC){
    if(userData[idP]>0.0){
        sx+=userData[idP];
        sy+=rat;
        sxy+=(userData[idP]*rat);
        sx2+=(userData[idP]*userData[idP]);
        sy2+=(rat*rat);
        n++;
    }
}
```

## 2.2. VARIABLES COSENO

```
if(idU==idC){
    lenY+=(rat*rat);
    Sxy+=(userData[idP]*rat);
}
```

## 2.3. VARIABLES MINKOWSKI

```
if(idU==idC){
    if(userData[idP]>0.0){
        Sxy+=pow((abs(userData[idP]-rat)),R);
        n++;
    }
}
```

Al final cuando cambia el Id del usuario procedemos a calcular las distancias según cada formula:

## 2.4. DISTANCIA PEARSON

```
num = (sxy-((sx*sy)/n));
den = sqrt(sx2-((sx*sx)/n))*(sqrt(sy2-((sy*sy)/n)));

if(den == 0){corr=-1e9;}
else{corr = double(num/den);}
```

## 2.5. DISTANCIA COSENO

```
cosD = Sxy/(sqrt(lenX)*sqrt(lenY));
```

## 2.6. DISTANCIA MINKOWSKI

```
dist = pow(Sxy,(1.0/R));
```

Al final verificamos en un vector de pares de tamaño K si la nueva distancia es mejor que el ultimo, si es así reemplazamos este con la nueva distancia y su Id de usuario y luego volvemos ordenar el vector para que siempre tengamos al mas lejano al final del vector.

```
if(k_near[kk-1].first < corr)
    k_near[kk-1]={corr,idC};
sort(k_near.begin(),k_near.end(),myComp1);
```

### 3. Lenguaje de programación librerías

El lenguaje de programación usado fue C++11, se utilizó también algunas librerías del lenguaje C para la lectura de archivos ya que tienen un mejor tiempo de respuesta en comparación con las librerías de C++, la razón de usar C++11 fue para la mejor manipulación de variable y mejor control de memoria. Usamos funciones de la librería STL como "sort" para el ordenamiento de los vectores. Otras librerías como `utility` ayudaron al almacenamiento de datos por su tipo de variable "pair". A parte implementamos nuestra propia función de búsqueda binaria para poder encontrar elementos dentro de un vector.

### 4. Proceso que sigue su sistema de recomendación

Lo primero que se hace es sacar los puntajes del usuario al que queremos recomendar para eso se llama a la función *find\_user\_data\_file*, seguidamente una vez obtenida la información se pasa a calcular el K-NN con la función *get\_kn* la cual llenara un vector donde se tendrá los ID's y las respectivas distancias de los K usuarios más cercanos. Una vez terminada esa función ordenamos el vector de los K más cercanos por sus ID's y pasamos a calcular las películas que recomendarían, para eso llamamos a la función *recomender* la cual calculara todas las películas cuyo puntaje pase el umbral dado. Para finalizar tendremos un vector donde estarán las ID's de las películas y su respectivo puntaje. Llamamos a la función *print\_product\_name* la cual se encargara de sacar los nombres de las películas seleccionadas previamente y las mostrara en pantalla. El vector *k\_near* tendrá los usuarios mas cercanos, al comienzo se llenara con valores lejanos.

```
int K=5;
double umbral=4;
k_near.assign(K,{-1e9,1});
int idQ = 16006;

find_user_data_file(idQ);
get_kn(idQ,K);

sort(k_near.begin(),k_near.end(),myComp2);

for(int i=0;i<k_near.size();i++){
    printf("%d%f\n",k_near[i].second,k_near[i].first);
}

recomender(K,umbral);

sort(recomP.begin(),recomP.end(),myComp2);

print_product_name();
```

## 5. Resolución de Preguntas

### 5.1. Recomendaciones

Usuario 4598 con TA = 0m24,682s, TP = 0m24,682s y RAM = 1.8 GB con similitud Coseno

Puntaje	Pelicula	Genero
4.5	Godfather, The (1972)	Crime—Drama
4.4	Clockwork Orange, A (1971)	Crime—Drama—Sci-Fi—Thriller
4.375	Lock, Stock & Two Smoking Barrels (1998)	Comedy—Crime—Thriller
4.375	Full Metal Jacket (1987)	Drama—War
4.375	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)	Mystery—Sci-Fi—Thriller
4.35714	Django Unchained (2012)	Action—Drama—Western
4.35714	Seven (a.k.a. Se7en) (1995)	Mystery—Thriller
4.3125	Shawshank Redemption, The (1994)	Crime—Drama
4.3	Schindler's List (1993)	Drama—War
4.28571	Trainspotting (1996)	Comedy—Crime—Drama
4.25	Gladiator (2000)	Action—Adventure—Drama
4.21429	Star Wars: Episode V - The Empire Strikes Back (1980)	Action—Adventure—Sci-Fi
4.21429	Shutter Island (2010)	Drama—Mystery—Thriller
4.2	Monty Python and the Holy Grail (1975)	Adventure—Comedy—Fantasy
4.2	One Flew Over the Cuckoo's Nest (1975)	Drama
4.16667	Dark Knight Rises, The (2012)	Action—Adventure—Crime—IMAX
4.16667	Usual Suspects, The (1995)	Crime—Mystery—Thriller
4.16667	Memento (2000)	Mystery—Thriller
4.14286	Star Wars: Episode IV - A New Hope (1977)	Action—Adventure—Sci-Fi
4.125	Fargo (1996)	Comedy—Crime—Drama—Thriller
4.125	Forrest Gump (1994)	Comedy—Drama—Romance—War
4.125	Godfather: Part II, The (1974)	Crime—Drama
4.125	Silence of the Lambs, The (1991)	Crime—Horror—Thriller
4.1	Wolf of Wall Street, The (2013)	Comedy—Crime—Drama
4.1	Goodfellas (1990)	Crime—Drama
4.1	Prestige, The (2006)	Drama—Mystery—Sci-Fi—Thriller
4.08333	Inglourious Basterds (2009)	Action—Drama—War
4	City of God (Cidade de Deus) (2002)	Action—Adventure—Crime—Drama—Thriller
4	Shining, The (1980)	Horror

Cuadro 1: Recomendaciones para el usuario 4598 con la distancia coseno con un K=10 y un umbral de puntaje de películas mayor a 4

Usuario 16006 con TA = 0m24,822s, TP = 0m24,822s y RAM = 1.8 GB con similitud Coseno

Puntaje	Pelicula	Genero
4.77778	Godfather, The (1972)	Crime—Drama
4.6875	Godfather: Part II, The (1974)	Crime—Drama
4.64286	Good, the Bad and the Ugly, The (Buono, il brutto, il cattivo, Il) (1966)	Action—Adventure—Western
4.625	Strangers on a Train (1951)	Crime—Drama—Film-Noir—Thriller
4.58333	Seven Samurai (Shichinin no samurai) (1954)	Action—Adventure—Drama
4.5	Vertigo (1958)	Drama—Mystery—Romance—Thriller
4.5	Mr. Smith Goes to Washington (1939)	Drama
4.5	To Kill a Mockingbird (1962)	Drama
4.5	Ran (1985)	Drama—War
4.5	My Life as a Dog (Mitt liv som hund) (1985)	Comedy—Drama
4.5	Simple Plan, A (1998)	Crime—Drama—Thriller
4.5	Amelie (Fabuleux destin d'Amélie Poulain, Le) (2001)	Comedy—Romance
4.5	After Hours (1985)	Comedy—Thriller
4.5	Interstellar (2014)	Sci-Fi—IMAX
4.5	Big Short, The (2015)	Drama
4.4375	Rear Window (1954)	Mystery—Thriller
4.41667	Blood Simple (1984)	Crime—Drama—Film-Noir
4.41667	Slumdog Millionaire (2008)	Crime—Drama—Romance
4.38889	Traffic (2000)	Crime—Drama—Thriller
4.375	Schindler's List (1993)	Drama—War
4.375	Exorcist, The (1973)	Horror—Mystery
4.375	Invasion of the Body Snatchers (1956)	Horror—Sci-Fi—Thriller
4.375	General, The (1926)	Comedy—War
4.375	Donnie Darko (2001)	Drama—Mystery—Sci-Fi—Thriller
4.375	Boyhood (2014)	Drama
4.375	Logan (2017)	Action—Sci-Fi
4.35714	12 Angry Men (1957)	Drama
4.35714	Lawrence of Arabia (1962)	Adventure—Drama—War
4.35714	Evil Dead II (Dead by Dawn) (1987)	Action—Comedy—Fantasy—Horror
4.33333	In the Name of the Father (1993)	Drama
4.33333	Bridge on the River Kwai, The (1957)	Adventure—Drama—War
4.33333	American Beauty (1999)	Drama—Romance
4.33333	American Graffiti (1973)	Comedy—Drama
4.33333	Marathon Man (1976)	Crime—Drama—Thriller
4.33333	Let the Right One In (Låt den rätte komma in) (2008)	Drama—Fantasy—Horror—Romance
4.3125	Dirty Dozen, The (1967)	Action—Drama—War
4.3125	Evil Dead, The (1981)	Fantasy—Horror—Thriller

Cuadro 2: Recomendaciones para el usuario 16006 con la distancia coseno con un K=10 y un umbral de puntaje de películas mayor a 4.3

Usuario 30503 con TA = 0m25,804s, TP = 0m25,804s y RAM = 2.0 GB con similitud Coseno

Puntaje	Pelicula	Genero
4.6875	Metropolis (1927)	Drama—Sci-Fi
4.5	Jetée, La (1962)	Romance—Sci-Fi
4.5	Tree of Life, The (2011)	Drama
4.45	Fargo (1996)	Comedy—Crime—Drama—Thriller
4.45	Star Wars: Episode V - The Empire Strikes Back (1980)	Action—Adventure—Sci-Fi
4.4444	Touch of Evil (1958)	Crime—Film-Noir—Thriller
4.41667	Boot, Das (Boat, The) (1981)	Action—Drama—War
4.41667	Rashomon (Rashômon) (1950)	Crime—Drama—Mystery
4.4	M (1931)	Crime—Film-Noir—Thriller
4.375	Hoop Dreams (1994)	Documentary
4.375	Manchurian Candidate, The (1962)	Crime—Thriller—War
4.35714	Raging Bull (1980)	Drama
4.35714	Bridge on the River Kwai, The (1957)	Adventure—Drama—War
4.35714	Chinatown (1974)	Crime—Film-Noir—Mystery—Thriller
4.35	Star Wars: Episode IV - A New Hope (1977)	Action—Adventure—Sci-Fi
4.35	Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)	Drama—Film-Noir—Romance
4.35	Die Hard (1988)	Action—Crime—Thriller
4.33333	Downfall (Untergang, Der) (2004)	Drama—War
4.3125	African Queen, The (1951)	Adventure—Comedy—Romance—War
4.3125	Platoon (1986)	Drama—War
4.3	Terminator 2: Judgment Day (1991)	Action—Sci-Fi
4.3	Crying Game, The (1992)	Drama—Romance—Thriller
4.3	Lives of Others, The (Das leben der Anderen) (2006)	Drama—Romance—Thriller
4.3	Departed, The (2006)	Crime—Drama—Thriller
4.3	There Will Be Blood (2007)	Drama—Western
4.28571	Night of the Living Dead (1968)	Horror—Sci-Fi—Thriller
4.28571	Grave of the Fireflies (Hotaru no haka) (1988)	Animation—Drama—War
4.27778	Once Upon a Time in the West (C'era una volta il West) (1968)	Action—Drama—Western
4.27778	Prestige, The (2006)	Drama—Mystery—Sci-Fi—Thriller
4.25	Taxi Driver (1976)	Crime—Drama—Thriller
4.25	Philadelphia Story, The (1940)	Comedy—Drama—Romance
4.25	Third Man, The (1949)	Film-Noir—Mystery—Thriller
4.25	Butch Cassidy and the Sundance Kid (1969)	Action—Western
4.25	Brick (2005)	Crime—Drama—Film-Noir—Mystery
4.25	Pan's Labyrinth (Laberinto del fauno, El) (2006)	Drama—Fantasy—Thriller
4.25	How the Grinch Stole Christmas! (1966)	Animation—Comedy—Fantasy—Musical
4.25	Planet Earth (2006)	Documentary
4.22222	Annie Hall (1977)	Comedy—Romance
4.22222	Let the Right One In (Låt den rätte komma in) (2008)	Drama—Fantasy—Horror—Romance
4.21429	Grand Budapest Hotel, The (2014)	Comedy—Drama

Cuadro 3: Recomendaciones para el usuario 30503 con la distancia coseno con un K=10 y un umbral de puntaje de películas mayor a 4.2

Usuario 283228 con TA = 0m34,791s, TP = 0m34,791s y RAM = 2.1 GB  
con similitud Coseno

Puntaje	Pelicula	Genero
4.92857	Godfather: Part II, The (1974)	Crime—Drama
4.7	Lady Vanishes, The (1938)	Drama—Mystery—Thriller
4.64286	Young Frankenstein (1974)	Comedy—Fantasy
4.625	Cinema Paradiso (Nuovo cinema Paradiso) (1989)	Drama
4.58333	Lady Eve, The (1941)	Comedy—Romance
4.58333	Double Indemnity (1944)	Crime—Drama—Film-Noir
4.4	Patton (1970)	Drama—War
4.4	Palm Beach Story, The (1942)	Comedy
4.4	Lord of the Rings: The Fellowship of the Ring, The (2001)	Adventure—Fantasy
4.35714	Treasure of the Sierra Madre, The (1948)	Action—Adventure—Drama—Western
4.25	Sting, The (1973)	Comedy—Crime
4.25	Strangers on a Train (1951)	Crime—Drama—Film-Noir—Thriller
4.25	Sullivan's Travels (1941)	Adventure—Comedy—Romance
4.25	Lord of the Rings: The Return of the King, The (2003)	Action—Adventure—Drama—Fantasy
4.1875	Chinatown (1974)	Crime—Film-Noir—Mystery—Thriller
4.1875	Back to the Future (1985)	Adventure—Comedy—Sci-Fi
4.1875	Shop Around the Corner, The (1940)	Comedy—Drama—Romance
4.16667	Goldfinger (1964)	Action—Adventure—Thriller
4.15	Indiana Jones and the Last Crusade (1989)	Action—Adventure
4.14286	Psycho (1960)	Crime—Horror
4.14286	Annie Hall (1977)	Comedy—Romance
4.14286	Groundhog Day (1993)	Comedy—Fantasy—Romance
4.125	Being John Malkovich (1999)	Comedy—Drama—Fantasy
4.125	Grapes of Wrath, The (1940)	Drama
4.11111	To Kill a Mockingbird (1962)	Drama
4.1	Miracle on 34th Street (1947)	Comedy—Drama
4.08333	High Noon (1952)	Drama—Western
4.08333	Shakespeare in Love (1998)	Comedy—Drama—Romance
4.07143	Clockwork Orange, A (1971)	Crime—Drama—Sci-Fi—Thriller
4.0625	Graduate, The (1967)	Comedy—Drama—Romance
4	Paths of Glory (1957)	Drama—War
4	Lawrence of Arabia (1962)	Adventure—Drama—War
4	Amadeus (1984)	Drama
4	Manchurian Candidate, The (1962)	Crime—Thriller—War
4	Ben-Hur (1959)	Action—Adventure—Drama
4	In the Heat of the Night (1967)	Drama—Mystery
4	Holiday Inn (1942)	Comedy—Musical
4	Memento (2000)	Mystery—Thriller
4	Lion in Winter, The (1968)	Drama

Cuadro 4: Recomendaciones para el usuario 283228 con la distancia coseno con un K=10 y un umbral de puntaje de películas mayor a 4

## 5.2. K mas cercanos

Correlación de Pearson para el usuario 16006

IdUsuario	Correlacion de Pearson
1277	1
1870	1
51639	1
54329	1
219635	1

Cuadro 5: Correlación de Pearson para el usuario 16006 con un tiempo de 0m24,354s, con un uso de RAM = 1.8 dando los 5 mas cercanos

Correlación de Coseno para el usuario 283228

IdUsuario	Correlacion de Coseno
8369	0.27393
61675	0.286801
90934	0.284435
112712	0.275811
140027	0.295898
147607	0.27612
173357	0.33241
220018	0.30609

Cuadro 6: Correlación de Coseno para el usuario 283228 con un tiempo de 0m23,657s, con un uso de RAM = 2 dando los 8 mas cercanos



Distancia de Manhattan para el usuario 283228

IdUsuario	Distancia de Manhattan
1	0
50	0
85	0
106	0
175	0
204	0
209	0
211	0

Cuadro 7: Distancia de Manhattan para el usuario 283228 con un tiempo de 0m24,075s, con un uso de RAM = 2 GB con un K = 8

Distancia Euclidiana para el usuario 30503

IdUsuario	Distancia Euclidiana
9	0
97	0
254	0
262	0
263	0
429	0
431	0
499	0
545	0
579	0

Cuadro 8: Distancia Euclidiana para el usuario 30503 con un tiempo de 0m17,314s, con un uso de RAM = 2 GB con un K =10

Correlación de Pearson para el usuario 4598

IdUsuario	Correlacion de Pearson
1878	1
2782	1
4792	1
4994	1
5972	1
7456	1
8982	1
9879	1
10021	1
158587	1

Cuadro 9: Correlación de Pearson para el usuario 4598 con un tiempo de 0m15,461s, con un uso de RAM = 1.8 GB con un K =10