**UNIVERSIDAD AUTONOMA DE MADRID**

**ESCUELA POLITECNICA SUPERIOR**

** **

**Grado en Ingeniería Informática**

**TRABAJO FIN DE GRADO**

**Framework orientado a algoritmos de recomendación basados en vecinos cercanos**

**Alejandro Gil Hernán**

**Tutor: Nombre y Apellidos**

**Ponente (si procede): Nombre y Apellidos**

**Febrero 2017**

**Framework orientado a algoritmos de recomendación basados en vecinos cercanos**

**AUTOR: Alejandro Gil Hernán**

**TUTOR: Nombre y Apellidos tutor/a**

**Grupo de la EPS (opcional)**

**Dpto. XXX**

**Escuela Politécnica Superior**

**Universidad Autónoma de Madrid**

**Febrero 2017**

**Resumen (castellano)**

Los sistemas de recomendación se han convertido en una pieza clave en internet, bien sea navegando a través de artículos científicos (Mendeley), música (Spotify, Last.fm), películas o vídeos (Netflix, IMDB, YouTube), personas (LinkedIn, Facebook, Twitter), así como plataformas online de venta de infinidad de productos (Amazon).

Dichos sistemas de recomendación están motivados a su vez por el crecimiento exponencial que ha experimentado la web en los últimos años y con la aparición de gran contenido digital.

Se caracterizan por observar la actividad de los usuarios y aprovecharla para predecir cuáles son los intereses de éstos, según los cuales se presentarán unos productos u otros, de una forma individual y personalizada.

En este TFG se explorará un tipo de algoritmo de recomendación muy habitual: vecinos cercanos. Estos algoritmos se suelen utilizar para hacer recomendaciones basadas en similitudes entre usuarios o entre objetos, siendo de esta forma un tipo de filtrado colaborativo; sin embargo, si la similitud tiene en cuenta atributos de los usuarios o de los objetos, también se podría utilizar para algoritmos basados en contenido. Este potencial permite, en principio, que se puedan utilizar en multitud de dominios, teniendo la ventaja adicional de que su salida es fácil de interpretar y analizar.

En este trabajo se plantea diseñar e implementar un framework orientado a generar y evaluar recomendaciones basadas en este tipo de algoritmos, así como el estudio sobre los distintos parámetros a configurar para discernir cuál es la combinación que aporta mejores resultados.

Por ello los objetivos principales serán obtener implementaciones generales, así como que se ejecuten de la forma más eficiente posible.

**Abstract (English)**

Recommender systems have become a key element in the internet in different aspects such as scientific articles browsing (Mendeley), music (Spotify, Last.fm), movies (Netflix, IMDB), people (LinkedIn, Facebook, Twitter), even online selling platforms (Amazon). Those recommender systems are also motivated by the exponential growth experimented by the web in the past years and with the appearance of substantial quantity of digital content.

are characterised by taking advantage of observing user´s activity to predict their interests. This results will be presented differently each time in an individual and personalised way.

In this bachelor, it will be explored a very common kind of recommendation algorithm: k-nearest neighbours (KNN). These algorithms are usually used to make recommendations based on similarities between users or items, being that way a kind of collaborative filtering; however, if the similarity considers user´s or items attributes, it also will be possible to use it for content-based algorithms. This potential allows to use them in many domains, having the advantage of the easily interpretable and analysable output.

This work contemplates to design and implement framework oriented to generate and evaluate recommendations based on these kind of algorithms, as well as the study of the different parameters configurations to distinguish which is the most beneficial combination.

Therefore, the principal objectives will be to obtain general implementations which will be executed in the most efficient way.

**Palabras clave (castellano)**

Sistema de recomendación, algoritmo, framework, vecinos cercanos (KNN), filtrado colaborativo, algoritmo basado en contenido.

**Keywords (inglés)**

Recommender system, algorithm, framework, nearest neighbours (KNN), collaborative filtering, content-based.

***Agradecimientos***

*Quiero agradecer en primer lugar a mi familia, sobre todo a mis padres y abuelos.*

*A mi madre por criarme y hacer de mí la persona que soy.*

*A mi padre por sus consejos.*

*A mis abuelos por tratarme como a un hijo.*

*A mi tutor Alejandro por guiarme en este trabajo y asignaturas de la carrera, por ser un profesor ejemplar con una dedicación como jamás he visto en el mundo docente.*

*Muchas gracias a mis amigos que son como hermanos y a Paula por estar siempre apoyándome.*

*Por último, dar las gracias a la Escuela Politécnica Superior por hacerme conocer a compañeros increíbles y formarme como profesional y como persona.*

**Alejandro Gil Hernán**

**ÍNDICE DE CONTENIDOS**

[**1. Introducción** 1](#_Toc175400053)

[1.1 Motivación 1](#_Toc175400054)

[1.2 Objetivos 1](#_Toc175400055)

[1.3 Organización de la memoria 1](#_Toc175400056)

[**2. Estado del arte** 3](#_Toc175400057)

[2.1 Subsección 3](#_Toc175400058)

[2.1.1 Subsubsección 3](#_Toc175400059)

[**3. Diseño** 5](#_Toc175400060)

[3.1 Subsección 5](#_Toc175400061)

[3.1.1 Subsubsección 5](#_Toc175400062)

[**4. Desarrollo** 7](#_Toc175400063)

[4.1 Subsección 7](#_Toc175400064)

[4.1.1 Subsubsección 7](#_Toc175400065)

[**5. Integración, pruebas y resultados** 9](#_Toc175400066)

[**6. Conclusiones y trabajo futuro** 9](#_Toc175400067)

[6.1 Conclusiones 9](#_Toc175400068)

[6.2 Trabajo futuro 9](#_Toc175400069)

[**Referencias** 11](#_Toc175400070)

[**Glosario** 13](#_Toc175400071)

[**Anexos**  I](#_Toc175400072)

[A Manual de instalación I](#_Toc175400073)

[B Manual del programador III](#_Toc175400074)

[C Anexo V](#_Toc175400075)

**INDICE DE FIGURAS**

[Figura 1: Logo EPS 3](#_Toc175399191)

**INDICE DE TABLAS**

# Introducción

Los sistemas de recomendación (SR) son herramientas software y técnicas que proveen al usuario de elementos que pueden resultarle interesantes y afines a sus gustos, facilitando la labor de búsqueda entre la inmensa cantidad de información de la que dispone. Éstas recomendaciones están relacionadas con las acciones que lleva a cabo el usuario: productos comprados, música que escucha, libros que lee… etc.

Hoy en día estamos más que acostumbrados a que prácticamente la totalidad de las aplicaciones informáticas de uso diario nos propongan contenido, ya sea porque Amazon nos recomiende un producto que comprar, Spotify una canción que escuchar o Youtube un vídeo que quizá nos guste. En definitiva, no nos resulta extraño que la propaganda *online* se adapte a nosotros con recomendaciones personalizadas.

## Motivación

Un punto importante, es el de la valoración y comparación de los resultados obtenidos en las recomendaciones, lo que se conoce como evaluación. Durante mucho tiempo la manera natural de evaluar tanto buscadores como recomendadores eran las métricas de error, pero es evidente que a lo largo de los años las tendencias varían, por ello, en la actualidad, están evolucionando hacia métricas basadas en ranking, ampliando un nuevo horizonte de estudio y desarrollo. Ahora queda pendiente la comparativa de efectividad entre los distintos algoritmos y variantes dentro de éstos con este nuevo tipo de métricas.

Este trabajo de fin de grado se ha realizado con la intención de crear un framework donde sea posible la comparación de resultados para recomendaciones generadas con un algoritmo basado en vecinos cercanos (KNN) y todas sus variantes posibles, ya que como se ha mencionado anteriormente, esto viene motivado por el cambio en las métricas de evaluación (la manera en que se mide la tasa de error, y por tanto, la eficacia en una recomendación).

Esto es importante ya que los sistemas de recomendación ofrecen una serie de ventajas a las plataformas de productos online:

* Incrementar el número de ítems vendidos: ya que a cada usuario se le muestra primero lo que potencialmente más le atrae.
* Aumentar la diversidad: Una característica es generar diversidad en las recomendaciones, sugiriendo al usuario ítems similares a sus gustos con un índice de popularidad menor.
* Incrementar la satisfacción del usuario.
* Incrementar la fidelidad del usuario.

Es por esto que el método de evaluación en este trabajo tendrá presente las tres métricas que creemos más influyentes actualmente: precisión, recall y NDCG (Normalized Cumulative Discounted Gain).

## Objetivos

La meta general de este trabajo es averiguar cuál de las configuraciones del algoritmo KNN basado en filtrado colaborativo implementadas arroja mejores resultados.

De forma particular, los objetivos a llevar a cabo son los siguientes:

* Fijar el punto de partida con un dataset concreto (no se va a tratar la recogida de información, por lo que se supone que la ésta se obtiene de forma externa al proyecto).
* Estudio de las variantes del algoritmo KNN así como de sus parámetros de entrada, con cada uno se realizará una variación
* Generar recomendaciones para los usuarios del sistema con la información del dataset inicial.
* Evaluar las recomendaciones generadas con las tres métricas más importantes.
* Sintetizar los resultados en una tabla comparativa, pudiendo escoger cualquier combinación deseada.

Para contrastar las diferentes variaciones simplemente será necesario buscar en la tabla comparativa de resultados la combinación deseada, que mostrará los valores de las distintas métricas de error.

## Organización de la memoria

La memoria consta de los siguientes capítulos:

* **…**

# Estado del arte

En la última década los sistemas de recomendación están adquiriendo una importancia esencial debido al crecimiento del comercio electrónico y a la repercusión que la world wide web está teniendo en los consumidores.

Es común asociar el mundo de los SR al de buscadores, aunque en realidad son algo dispares, a diferencia de un buscador donde es necesario que el usuario introduzca explícitamente lo que quiere buscar, un recomendador actúa como una entidad activa, recuperando información de manera implícita de los usuarios a través de sus interacciones y sugiriendo ítems de una manera transparente.

A su vez, los SR comparten cualidades con el aprendizaje automático, el sistema también aprende los intereses y gustos del usuario detectando patrones de comportamiento. Varias técnicas como el cálculo de vecino, factorización de matrices o las particiones del dataset son aspectos que comparten ambos campos.

Los usuarios de un SR pueden tener características muy distintas, para poder realizar recomendaciones personalizadas, lo esencial es estudiar la estructura que tiene la información que se posee. Ésta información puede estar estructurada de varias maneras, la selección de que atributos son relevantes depende de la técnica de recomendación que se vaya a utilizar. Por ejemplo, en filtrado colaborativo, los usuarios son estructurados como una lista que contiene los diferentes ítems que ese usuario ha puntuado, aunque también pueden ser descritos por su patrón de comportamiento, como las acciones que llevan a cabo a la hora de navegar, por ejemplo, donde se hace click, de forma que es posible conocer los intereses también de esta manera.

Aun así, las puntuaciones o *ratings* son la forma más común de recoger los datos en un SR. Pueden ser obtenidos de manera implícita (el número de veces que se escucha una canción) o explícita (la puntuación numérica que un usuario concreto da a una película).

Existen diferentes tipos de *ratings* explícitos:

* Numéricos: puntuación normalmente comprendida entre 1 y 5.
* Binarios: me gusta, no me gusta
* Ordinales (en desacuerdo, de acuerdo, muy de acuerdo…).

Para los *ratings* implícitos depende de la plataforma donde esté implementada el SR, en el caso de canciones, las veces que ha sido escuchada, para productos, simplemente haber hecho click, para vídeos el tiempo de visualización… etc. Con este tipo de *ratings*, el hecho de que un usuario no haya interaccionado con un ítem concreto, puede indicar que no le interesa, es decir, que posee una puntuación baja, por otro lado, la ausencia de *ratings* explícitos no aporta información a la hora de recomendar.

Un evento a destacar ocurrido en los últimos años ha sido el *Netflix prize* (premio Netflix), dónde la famosa empresa multinacional de servicios multimedia (series y películas)

The Netflix dataset contains more than 100 million date- stamped movie ratings

The Netflix Prize was an open competition for the best [collaborative filtering](https://en.wikipedia.org/wiki/Collaborative_filtering) [algorithm](https://en.wikipedia.org/wiki/Algorithm) to predict user ratings for [films](https://en.wikipedia.org/wiki/Film), based on previous ratings without any other information about the users or films, i.e. without the users or the films being identified except by numbers assigned for the contest.

The competition was held by [Netflix](https://en.wikipedia.org/wiki/Netflix), an online [DVD](https://en.wikipedia.org/wiki/DVD)-rental and video streaming service, and was open to anyone not connected with Netflix (current and former employees, agents, close relatives of Netflix employees, etc.)

On 21 September 2009, the grand prize of US$1,000,000 was given to the [BellKor's Pragmatic Chaos](https://en.wikipedia.org/w/index.php?title=BellKor%27s_Pragmatic_Chaos&action=edit&redlink=1) team ([link](http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf)) which bested Netflix's own algorithm for predicting ratings by 10.06%

## Tipos de algoritmos

### Subsubsección

# Diseño

## Subsección

### Subsubsección

# Desarrollo

## Subsección

### Subsubsección

# Integración, pruebas y resultados

# Conclusiones y trabajo futuro

## Conclusiones

## Trabajo futuro

# Referencias

1. En las referencias figurarán los autores (opcionalmebte los editors), el título del artículo, el nombre de la revista o libro, el volumen y número de la revista, las páginas del artículo, la fecha de edición,. A continuación se listan algunos ejemplos
2. K.N. Platanioitis, C.S. Regazzoni (eds.), “Special Issue in Visual-centric Surveillance Networks and Services”, IEEE Signal Processing Magazine, 22(2), Marzo 2005.
3. B.S. Manjunath, P. Salembier, T. Sikora (eds.), “Introduction to MPEG 7: Multimedia Content Description Language,”, John Wiley and Sons, 2002
4. G. R. Bradski, “Computer vision face tracking as a component of a perceptual user interface,” en Proc.IEEE Workshop on Applications of Computer Vision, Princeton, NJ, October 1998, pp. 214–219.
5. A. D. Bue, D. Comaniciu, V. Ramesh, and C. Regazzoni, “Smart cameras with real-time video object generation,” in Proc. IEEE Intl. Conf. on Image Processing, Rochester, NY, volume III, 2002, pp. 429–432.
6. P. Anandan. “A computacional cuadrowork and an algorithm for the measurement of visual motion”, International Journal of Computer Vision, 2(3):283-310, January, 1989.
7. W.J. Ruckelidge. “Efficient Computation of the minimum Hausdorff Distance for Visual Recognition”, Phd thesis, Cornell Universitym 1995. CS-TR1454
8. “Extensible Markup Language (XML) 1.0 (Second Edition)”, W3C Recommendation 6 October 2000 <http://www.w3.org/TR/REC-xml>
9. William H. Press, Saul A.Teukolsky, William T. Vetterling, Brian P. Flannery. “Numerical Recipes in C – The art of Scientific Computing 2nd Edition”. Cambridge University Press

# Glosario

API Application Programming Interface

# Anexos

## Manual de instalación

## Manual del programador

## Anexo …

## 