**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 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

A lo largo de los últimos años, han surgido diferentes algoritmos y técnicas de recomendación las cuales eran evaluadas con las mismas métricas.

Recientemente, se están revisando estos métodos de evaluación para medir la eficacia de los Sistemas de Recomendación (SR), ampliando un nuevo horizonte de estudio y desarrollo de algoritmos más óptimos y eficaces. Ahora queda pendiente la comparativa de efectividad entre los distintos algoritmos y variantes dentro de éstos, dado que los que funcionaban bien en un pasado no tienen por qué hacerlo ahora.

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 cercanas (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 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 información se obtiene de forma externa a este proyecto).
* Estudio de las variantes del algoritmo KNN así como de sus parámetros de entrada, con cada uno se realizará una variación
* recomendar
* evaluar
* síntesis de resultados
* simplemente será necesario buscar en la tabla comparativa de resultados la combinación deseada, que devolverá los valores de las distintas métricas de error.

## Organización de la memoria

La memoria consta de los siguientes capítulos:

* **…**

# Estado del arte

## Subsección

### Subsubsección

# Diseño

## Subsección

### Subsubsección

Users of a RS, as mentioned above, may have very diverse goals and characteristics.

In order to personalize the recommendations and the human-computer

interaction, RSs exploit a range of information about the users. This information

can be structured in various ways and again the selection of what information to

model depends on the recommendation technique.

For instance, in collaborative filtering, users are modeled as a simple list containing

the ratings provided by the user for some items. In a demographic RS, sociodemographic

attributes such as age, gender, profession, and education, are used.

User data is said to constitute the user model [21, 32]

Users can also be described by their behavior pattern data, for example, site

browsing patterns (in a Web-based recommender system) [107], or travel search

patterns (in a travel recommender system) [60]. Moreover, user data may include relations

between users such as the trust level of these relations between users (Chapter

20). A RS might utilize this information to recommend items to users that were

preferred by similar or trusted users.

In fact, ratings are the most popular form of transaction data that a RS collects.

These ratings may be collected explicitly or implicitly. In the explicit collection of

ratings, the user is asked to provide her opinion about an item on a rating scale.

According to [93], ratings can take on a variety of forms:

*•* Numerical ratings such as the 1-5 stars provided in the book recommender associated

with Amazon.com.

*•* Ordinal ratings, such as “strongly agree, agree, neutral, disagree, strongly disagree”

where the user is asked to select the term that best indicates her opinion

regarding an item (usually via questionnaire).

*•* Binary ratings that model choices in which the user is simply asked to decide if

a certain item is good or bad.

*•* Unary ratings can indicate that a user has observed or purchased an item, or

otherwise rated the item positively. In such cases, the absence of a rating indicates

that we have no information relating the user to the item (perhaps she purchased

the item somewhere else).

Another form of user evaluation consists of tags associated by the user with the

items the system presents

Amazon.com she will be provided with a long list of books. In return, the user may

click on a certain book on the list in order to receive additional information. At this

point, the system may infer that the user is somewhat interested in that book.

Content-based: The system learns to recommend items that are similar to the

ones that the user liked in the past. The similarity of items is calculated based on the

features associated with the compared items. For example, if a user has positively

rated a movie that belongs to the comedy genre, then the system can learn to recommend

other movies from this genre.

Collaborative filtering: The simplest and original implementation of this approach

[93] recommends to the active user the items that other users with similar

tastes liked in the past. The similarity in taste of two users is calculated based on

12 Francesco Ricci, Lior Rokach and Bracha Shapira

the similarity in the rating history of the users. This is the reason why [94] refers

to collaborative filtering as “people-to-people correlation.” Collaborative filtering is

considered to be the most popular and widely implemented technique in RS.

# 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 …

## 