**UNIVERSIDAD AUTÓNOMA DE MADRID**

**ESCUELA POLITÉCNICA SUPERIOR**

** **

**Grado en Ingeniería Informática**

**TRABAJO FIN DE GRADO**

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

**Alejandro Gil Hernán**

**Tutor: Alejandro Bellogín Kouki**

**Ponente: Pablo Castells Azpilicueta**

**Febrero 2017**

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

**AUTOR: Alejandro Gil Hernán**

**TUTOR: Alejandro Bellogín Kouki**

**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 y 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 and they 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 thesis, 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

1.1 Motivación 1

1.2 Objetivos 2

1.3 Organización de la memoria 3

2. Estado del arte 2

2.1 Tipos de algoritmos 3

2.1.1 Algoritmos basados en contenido 3

2.1.2 Algoritmos basados en filtrado colaborativo 3

2.1.3 Algoritmos generales 6

2.2 Normalización de ratings 6

2.3 Selección de vecinos 7

2.4 Comparación entre User KNN e Item KNN 8

2.5 Métricas de evaluación 9

3. Diseño y desarrollo 13

3.1 Descripción del sistema 13

3.2 Requisitos del sistema 14

3.2.1 Requisitos funcionales 14

3.2.2 Requisitos no funcionales 14

3.3 Diseño 15

3.4 Desarrollo y codificación 15

3.4.1 Librerías externas utilizadas 17

3.4.2 Cálculo aproximado de vecinos 17

3.4.3 Annoy vs NMSLIB 24

3.4.4 Apache Thrift y NMSLIB 25

4. Integración, pruebas y resultados 27

5. Conclusiones y trabajo futuro 27

Referencias 29

Glosario 31

Anexos I

A Manual de instalación I

Manual del programador III

B Anexo … - 1 -

**INDICE DE FIGURAS**

[Figura 1. Valor de las ventas del comercio electrónico en EEUU 1](#_Toc471825593)

[Figura 2. Elección de vecinos 8](#_Toc471825594)

[Figura 3. Conjunto intersección entre ítems recomendados y relevantes 10](#_Toc471825595)

[Figura 4. Precisión frente a recall 10](#_Toc471825596)

[Figura 5. Esquema de arquitectura. 13](#_Toc471825597)

[Figura 6. Esquema de las variantes en kNN 15](#_Toc471825598)

[Figura 7. Mapa bidimensional de usuarios 18](#_Toc471825599)

[Figura 8. Mapa de distribución de usuarios en el espacio 18](#_Toc471825600)

[Figura 9. Mapa con tres particiones aleatorias 19](#_Toc471825601)

[Figura 10. Árbol resultante de realizar tres particiones 19](#_Toc471825602)

[Figura 11. Mapa de particiones para k=10 20](#_Toc471825603)

[Figura 12. Árbol binario de particiones para k=10. 20](#_Toc471825604)

[Figura 13. Búsqueda de un usuario en el espacio 21](#_Toc471825605)

[Figura 14. Árbol resultante de la búsqueda de un usuario 21](#_Toc471825606)

[Figura 15. Mapa de ampliación en la búsqueda de vecinos 22](#_Toc471825607)

[Figura 16. Árbol de ampliación en la búsqueda de vecinos 22](#_Toc471825608)

[Figura 17. Vecindario de candidatos a vecinos más próximos en el espacio 23](#_Toc471825609)

[Figura 18. Radio que comprende los vecinos más próximos en el espacio 23](#_Toc471825610)

[Figura 19. Mapa bidimensional de vecinos con particiones aleatorias 24](#_Toc471825611)

[Figura 20. Annoy vs NMSLIB 24](#_Toc471825612)

[Figura 21. Esquema de empleo de NMSLIB 25](#_Toc471825613)

**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

Los sistemas de recomendación nacen con el objetivo de facilitar la toma de decisiones al usuario, así como de aumentar el beneficio de empresas dedicadas al comercio *online*. Desde la aparición de la WWW (*World Wide Web*) el comercio en internet ha experimentado un crecimiento exponencial, hasta el punto de que la comisión nacional de los mercados y la competencia (CNMC) determinó que la facturación de las tiendas *online* en España crecía a un 25% interanual.



Figura . Valor de las ventas del comercio electrónico en EEUU

Para conseguir estos buenos resultados, los puntos más importantes de los SR son 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 estos sistemas 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 (*normalised Discounted Cumulative Gain*).

## Objetivos

La meta general de este trabajo es averiguar cuál de las configuraciones implementadas del algoritmo KNN basado en filtrado colaborativo 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 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.
* Estudio de métodos alternativos para el cálculo de vecinos, sin basarse en el procedimiento tradicional de los SR para analizar su efecto en la eficiencia.
* Generar recomendaciones para los usuarios del sistema con la información del dataset inicial.
* Evaluar las recomendaciones generadas con las tres métricas más importantes mencionadas anteriormente.
* Sintetizar los resultados en una tabla comparativa, pudiendo escoger cualquier combinación deseada.

Para contrastar las diferentes variaciones se realizarán experimentos exhaustivos donde se podrán comparar los valores de las distintas métricas obtenidos por cada variante, lo cual nos permitirá concluir qué alternativas son mejores en cada caso.

## 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 Internet está teniendo en los consumidores.

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

A su vez, los SR comparten cualidades con el aprendizaje automático, ya que el sistema también aprende los intereses y gustos del usuario detectando patrones de comportamiento. Varias técnicas de los SR como el cálculo de vecinos, la factorización de matrices o las particiones del dataset utilizan conceptos y herramientas conocidas en ambos campos.

Los usuarios de un SR pueden tener características muy distintas; no obstante, 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 qué 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, dónde 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 da a un ítem concreto).

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…).

Por otro lado, los tipos de ratingsimplícitos dependen de la plataforma donde esté implementado el SR, por ejemplo, en el caso de canciones, el rating representa las veces que ha sido escuchada cada canción o cada artista; para productos, puede representar simplemente el haber hecho click, y 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 normalmente se entiende como que no conoce dicho ítem, mientras que, si el valor es muy bajo, se interpreta como que tiene poco interés en el mismo. En cambio, la ausencia de ratings explícitos aporta un matiz diferente a la hora de recomendar, ya que el usuario puede conocer el ítem, pero prefiere no invertir ese tiempo en dar dicha información al sistema.

Un evento a destacar ocurrido en los últimos años ha sido el *Netflix prize* (premio Netflix) [10], donde la famosa empresa multinacional de servicios multimedia (series y películas) propuso una competición abierta en busca del mejor algoritmo de filtrado colaborativo para generar recomendaciones de películas. En el reto podía participar cualquier persona excepto aquellas cercanas a la propia empresa.

El 21 de septiembre de 2009 el gran premio de US$1.000.000 fue otorgado al equipo de BellKor's Pragmatic Chaos, formado por la unión de los equipos Bellkor in BigChaos and Pragmatic Theory, quienes consiguieron una mejora de 10,06% sobre el algoritmo de Netflix.

Este reto es importante ya que junto a él surge la motivación de mejorar y optimizar los métodos de recomendación, así como la publicación de un dataset muy amplio y relevante a nivel mundial con 480.000 usuarios, 17.000 películas y más de 100 millones de ratings.

## Tipos de algoritmos

Dependiendo de la forma en que los algoritmos tengan en cuenta los atributos de los ítems o usuarios, podemos concluir que existen dos tipos de algoritmos principales: basados en contenido y basados en filtrado colaborativo.

### Algoritmos basados en contenido

Cuando se aplican métodos basados en contenido, el sistema aprende a recomendar ítems similares a los que le han gustado al usuario en un pasado [6]. Esta similitud es calculada basándose en las características asociadas a los ítems. En el caso de que un usuario haya puntuado con buena nota una película cuyo género es comedia, es muy probable que se generen recomendaciones de películas con este mismo género. En el caso de ítems más complejos como documentos de texto, se calcula el *TF-IDF* (Term Frequency-Inverse Document Frequency) que comprende las palabras o términos más relevantes.

Los algoritmos más comunes de éste género son *Rocchio* y KNN.

### Algoritmos basados en filtrado colaborativo

A diferencia de los métodos basados en contenido, el filtrado colaborativo se basa en los ratings de los usuarios del sistema. La idea clave es que, si dos usuarios *u* y *v* son similares, el sistema recomendará a *u*, ítems que le hayan gustado a *v*, ya que en teoría sus gustos son similares.

El filtrado colaborativo soluciona algunas de las limitaciones de los algoritmos basados en contenido. Por ejemplo, los ítems cuyo contenido es difícil de obtener pueden seguir siendo recomendados a usuarios por la similitud con otros. Otra ventaja es que el sistema puede recomendar mucha más variedad de elementos, nunca produciéndose el encasillamiento que ocurre con los basados en contenido, ya que las tendencias de una comunidad de usuarios pueden ir variando, a la vez que el SR se adapta a ellas.

***Basados en usuario (User KNN)***

Se recomiendan al usuario los ítems que han gustado a usuarios similares a éste.



***Basado en ítem (Item KNN)***

Se recomiendan al usuario los ítems que se parecen a ítems que le han gustado.



Los algoritmos KNN requieren de un vecindario para realizar el cálculo de ratings [7]. Si se aplica un filtrado colaborativo basado en usuario, el vecindario se compondrá de un conjunto de usuarios ordenados por similitud, en orden decreciente, tomando como referencia el usuario al que se le quiere recomendar. En el caso de utilizar filtrado colaborativo basado en ítem, la estructura es la misma, teniendo ítems en vez de usuarios en el vecindario.

Dichas similitudes representan el factor de semejanza y se calculan todas para cada par de usuarios o ítems. Existen diferentes métodos para medirlas, en este trabajo se han estudiado las siguientes similitudes:

* **Coseno**: Quizá sea la más famosa de todas, su función es medir el ángulo entre los vectores (ratings) para cada par de usuarios o ítems. Si ambos vectores apuntan al mismo lugar (son iguales) el valor de la similitud tendrá como valor 1, en el caso contrario que sean ortogonales el coseno se anula, devolviendo una similitud de 0.

****

En el caso de coseno basado en ítem:



* **Pearson**: Muy similar al coseno, con la diferencia de que Pearson tiene en cuenta la puntuación media del usuario o ítem para suavizar las puntuaciones con valores extremos.

****

Para ítems:



Se han implementado dos variantes de ésta similitud, ya que a la hora de realizar el cálculo se pueden tomar todos los usuarios o ítems del dataset o simplemente aquellos que posean ratings en común.

* **Jaccard**: Mide el grado de similitud entre dos conjuntos como la cardinalidad de la intersección de ambos conjuntos dividida por la cardinalidad de su unión. Devuelve 0 si los conjuntos no poseen ningún valor en común, tendiendo a 1 a medida que aumenta el número de elementos compartidos:

****

***Factorización de matrices***

Existen otros métodos no estudiados en este trabajo, como los algoritmos de factorización de matrices, que son bastante exitosos. La mayoría de estos métodos caracterizan a los usuarios y a los ítems como vectores derivados de los patrones de ratings, dividiendo la matriz inicial en productos de diferentes matrices. Éstos métodos se han popularizado recientemente por tener un buen equilibrio entre escalabilidad y precisión en las recomendaciones, también son más flexibles a la hora de modelar situaciones del mundo real. Por norma general, los datos de entrada más convenientes para este tipo de algoritmos son ratings explícitos, pero en caso de no tenerlos, es posible añadir información adicional, utilizando ratings implícitos que reflejan indirectamente la opinión de los usuarios simplemente observando su comportamiento.

Hay diferentes variantes de algoritmos para obtener la factorización de matrices como pLSA (probabilistic Latent Semantic Analysis), SVD (Singular Value Decomposition), SVDN (SVD no-empty entries), HSVDN (SVD with Hypergraph transformation).

### Algoritmos generales

Los algoritmos base o generales (*baseline*) no ofrecen recomendaciones personalizadas, pues no tienen en cuenta ningún dato acerca de los usuarios.

***Popularidad***

Recomienda los ítems más populares, todos los usuarios tienen el mismo ranking. La popularidad implica el número de interacciones totales con un ítem, ya sean reproducciones de un vídeo o canción, productos más vendidos o los ratings más altos.

Éste algoritmo es de los que más se usan a nivel comercial, sin necesidad de utilizar un sistema de recomendación, debido a su fácil implementación y aporte de información útil para el usuario.

***Random***

Recomienda ítems de manera aleatoria, representa la probabilidad de escoger un ítem en un conjunto de datos, por lo que cuanto más denso sea, más probabilidad de recomendar al usuario algo que le guste.

## Normalización de ratings

La normalización se aplica para acotar el valor de un rating a una escala deseada o para tener (o no tener) en cuenta ciertas desviaciones en los datos [7]. Existen las siguientes variantes:

* **Mean-Centering**: se compara el rating calculado con la media del usuario objetivo. Así se puede saber la apreciación (positiva o negativa) de un cierto usuario por un ítem observando el rating normalizado.



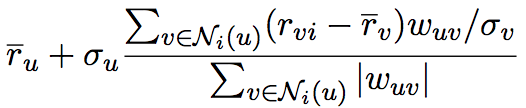
Para ítems:



* **Z-score**: supóngase el caso donde se tienen dos usuarios: A y B, el primero de ellos, posee una dispersión de ratings alta (varían mucho), mientras que el segundo siempre puntúa con un rating igual a 3.

El hecho de que B puntúe un ítem con un rating de 5 implica más apreciación por parte de ese usuario que si dicha puntuación se la hubiera dado A. En mean-centering se eliminan estas desviaciones tomando únicamente la media.

Esta apreciación es la que mide z-score dividiendo el mean-centering del usuario entre la desviación típica de los ratings:



Para ítems:



## Selección de vecinos

El número de vecinos próximos y el criterio utilizado puede tener un gran impacto en la calidad del sistema de recomendación [7]. La selección de los vecinos se lleva a cabo en dos pasos:

* Filtrado global donde solamente los candidatos más afines son seleccionados.
* Una fase de predicción donde se escogen los posibles mejores candidatos.

***Filtrado global de vecinos***

* **Top-N**: Para cada usuario, se guardan los N vecinos más cercanos con sus respectivas similitudes. Es necesario escoger de manera cautelosa este N por motivos de rendimiento y precisión. Si N es muy grande, se necesitará gran cantidad de memoria para almacenar todo el vecindario con los datos asociados y la predicción será lenta. Sin embargo, escogiendo un N muy pequeño es probable que algunos ítems nunca sean recomendados debido a la baja cobertura.



Figura . Elección de vecinos

* **Threshold**: En vez de mantener un número fijo de vecinos (N), threshold establece un umbral, el cual deben superar los valores de las similitudes.

Ésta técnica es algo más flexible que la anterior, pues solo los vecinos más significativos son escogidos, aun así, dicho umbral es más complicado de establecer.

* **Negative filtering**: Por lo general, las correlaciones negativas entre usuarios son menos fiables que las positivas, esto se debe intuitivamente a que una correlación positiva entre dos usuarios es un buen indicador de su correspondencia a un grupo determinado (adolescente, fans de un género determinado…).

Aun así, una correlación negativa puede implicar pertenencia a grupos opuestos, y dependiendo del conjunto de datos en el que se esté trabajando, se pueden ignorar.

## Comparación entre User KNN e Item KNN

***Precisión***

Depende del ratio entre el número de usuarios y los ítems del sistema, influyendo también si los ratings están distribuidos uniformemente sobre los ítems o no. En general, es preferible un número pequeño de vecinos fiables a una gran cantidad de vecinos con una similitud no tan alta.

***Eficiencia***

Tanto la memoria utilizada por un recomendador como el coste computacional dependen del ratio usuarios/ítems.

Por lo tanto, cuando el número de usuarios sobrepasa al número de ítems, correspondiendo al caso normal, la recomendación basada en ítem requiere mucha menos memoria y coste, ya que se tienen que calcular muchas menos similitudes. Aun así, el tiempo necesario para calcular una recomendación en la fase online*,* que depende solamente del nº de ítems relevantes y el nº máximo de vecinos es el mismo para los dos algoritmos.

En la práctica, el coste necesario para computar las similitudes es mucho menos costoso que el peor caso (teóricamente hablando) dado que los usuarios puntúan unos pocos ítems, y sólo se calculan las similitudes entre dos usuarios que hayan puntuado ítems en común, por lo que el conjunto de vecinos candidatos suele ser pequeño.

***Estabilidad***

La elección entre estos dos algoritmos se puede basar en la frecuencia y cantidad de cambios en los usuarios e ítems del sistema. Si la lista de ítems es bastante estática en comparación a la de usuarios, una recomendación basada en ítem es preferible dado que las similitudes entre ítems se pueden calcular en intervalos de tiempo separados, pudiendo de esta manera realizar recomendaciones a usuarios nuevos.

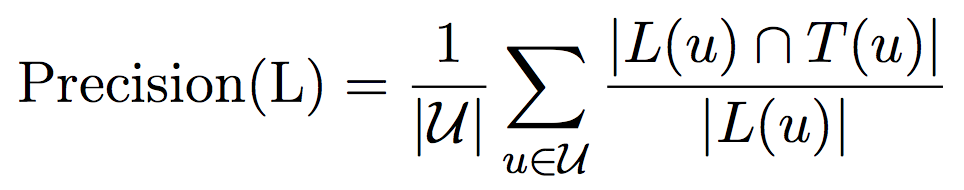
Por el contrario, en sistemas donde el número de ítems cambia constantemente, los métodos basados en usuarios prueban ser más estables.

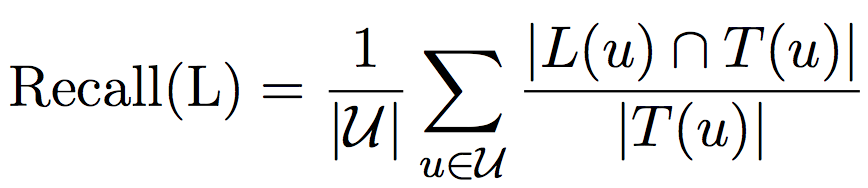
Un aspecto importante que diferencia a estos algoritmos es por ejemplo la variedad:

mientras que una recomendación basada en ítem tendrá poca variedad, pues siempre recomendará ítems parecidos que le hayan gustado al usuario, provocando cierto estancamiento, por otra parte, una recomendación basada en usuario escogerá ítems algo más distantes a los gustos iniciales del usuario, que en ciertos casos es preferible al “estancamiento” antes mencionado.

## Métricas de evaluación

Tradicionalmente, la eficacia de las recomendaciones se ha calculado mediante métricas de error como MAE (Mean Squared Error) o RMSE (Root Mean Squared Error) que representan la distancia entre la puntuación generada por el recomendador y la que realmente hizo el usuario. El problema reside en que este tipo de métricas están orientadas a la predicción de ratings, y éstos, no están siempre disponibles, ya que es bastante probable que existan ítems que el usuario no ha puntuado. Existen a su vez métricas que tienen en cuenta la novedad y diversidad de los ítems, no estudiadas en este trabajo. Por ello, el foco de trabajo se ha dirigido hacia tres métricas: precisión, recall y nDCG (*normalised Discounted Cumulative Gain*).





La métrica de *precisión* obtiene el conjunto intersección entre los ítems recomendados (*L*(*u*)) y los relevantes (*T*(*u*)) y dividiéndolo entre el conjunto total de recomendados. *Recall* por el contrario, divide la misma intersección por el total de ítems relevantes, éstos son los que el usuario ha puntuado, mientras que los recomendados hacen referencia a los que el sistema ha proporcionado para un usuario concreto.

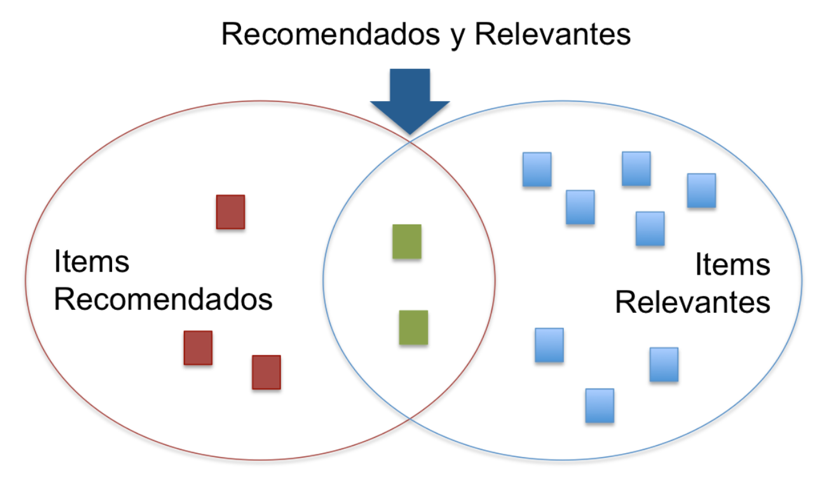


Figura . Conjunto intersección entre ítems recomendados y relevantes

Al aumentar el recall (la proporción de elementos relevantes) disminuye la precisión, ya que, si hay un número alto de relevantes, es menos probable llegar a recomendarlos todos, de esta manera, hay un compromiso entre ambas métricas.

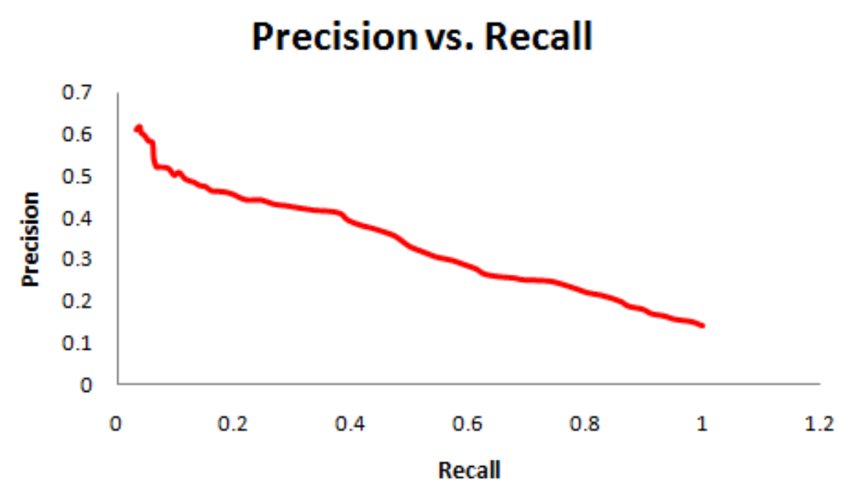
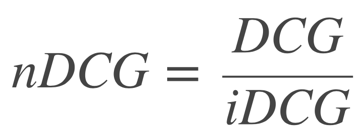
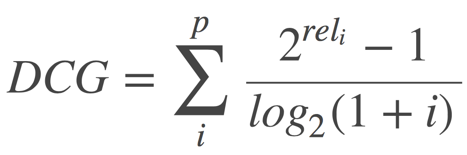


Figura . Precisión frente a recall

La última métrica por tratar (nDCG) se basa en la relevancia de los ítems teniendo en cuenta su posición en el ranking de recomendación, de modo que el valor de esta métrica será alto si los ítems más relevantes son los primeros posicionados.





Primero se calcula el *DCG* (*Discounted Cummulative Gain*), que representa la suma de grados de relevancia penalizados por lo tarde que aparezcan en el ranking, como se menciona anteriormente. *iDCG* hace referencia al valor ideal de *DCG*, es decir, que los n ítems escogidos estuvieran en las n primeras posiciones.

# Diseño y desarrollo

En este capítulo se describen los aspectos técnicos del proyecto en cuanto a desarrollo de software, incluyendo la integración de librerías externas, amplio rango de pruebas con su correspondiente análisis para poder comparar el rendimiento de las diferentes versiones del algoritmo.

De manera general, se puede subdividir el proyecto en módulos diferenciándolos por funcionalidad:

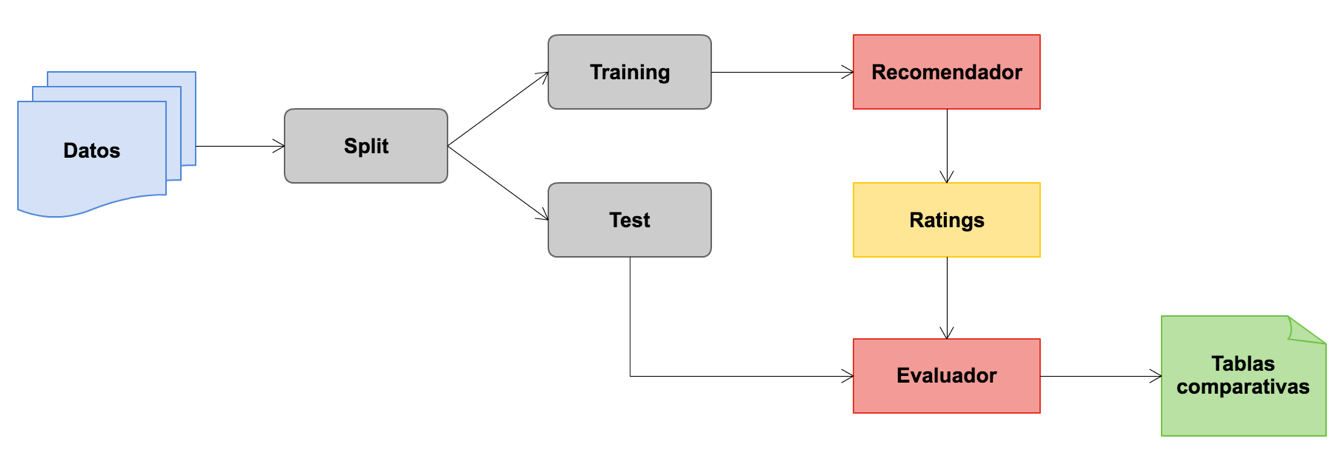


Figura . Esquema de arquitectura.

Todo parte de un conjunto de datos (*dataset*), que contiene los ficheros de usuarios, ítems, ratings... con sus correspondientes particiones de entrenamiento y test (*splits*). Este trabajo no se focaliza en la minería de datos, por lo que tomaremos dichos ficheros ajenos al sistema. Tampoco se realizan los splits del dataset ya que vienen incluidos.

Seguidamente, tenemos los dos grandes grupos de evaluación y recomendación, los cuales albergan el grueso del trabajo. Mientras que el recomendador se encarga de la generación de ficheros con las recomendaciones personalizadas de los usuarios a partir de cualquier combinación deseada, el evaluador toma esos ficheros para calcular la eficacia del algoritmo con la configuración de entrada. Este trabajo contiene todas las combinaciones posibles que permite el algoritmo para las parametrizaciones que han sido implementadas.

Una vez evaluadas las recomendaciones, ya disponemos de todos los datos necesarios para ensamblar nuestro framework, que serán extrapolados en formato de tablas para poder discernir la mejor solución con un simple vistazo.

## Descripción del sistema

Teniendo en cuenta la funcionalidad que iba a adquirir el proyecto, así como los algoritmos que se iban a implementar, se optó por el uso de Java como lenguaje de programación al tener ventajas como:

* Lenguaje de alto nivel orientado a objetos.
* Gran soporte de librerías.
* Multiplataforma.
* Lenguaje conocido. Al ser un lenguaje altamente usado en la carrera, no requería un elevado esfuerzo de aprendizaje más allá de las novedades introducidas en la versión 1.8 (programación funcional), versión en la que se basa la principal librería del proyecto.

## Requisitos del sistema

### Requisitos funcionales

* + - 1. Permitir predicciones de rating basadas en usuario e ítem.
      2. A la hora generar recomendaciones el sistema debe permitir la variación de distintos parámetros como:
* Elegir la similitud del vecindario (coseno, jaccard o pearson (con sus dos variantes).
* Elegir la similitud entre los usuarios o ítems en el momento de calcular el rating, puede ser la misma que la del vecindario.
* Permitir la normalización de la similitud.
* Permitir la normalización de los ratings (*Standard* (sin normalizar), *Mean-Centering* o *Z-Score).*
* Escoger el cálculo de vecinos mediante similitud o NMSLIB.
* Permitir aplicar un umbral a la similitud (threshold).
  + - 1. Si ocurre algún fallo interno, el sistema informará del error ocasionado.
      2. El sistema no dispondrá de interfaz gráfica.
      3. El sistema generará ficheros de texto plano con los resultados de los experimentos.

### Requisitos no funcionales

1. El sistema debe poder ser accedido desde un repositorio GIT (p.ej GitHub1).
2. El sistema funcionará en cualquier plataforma con versión de Java 1.8 a superior.

## Diseño

Diagrama de clases

## Desarrollo y codificación

El primer paso en el desarrollo es el estudio y la integración de la librería RankSys. De ella se han utilizado algunas funcionalidades como base para el proyecto principal como:

* Similitudes de coseno tradicional y Jaccard.
* Módulo de carga de datos del dataset.
* Estructuras de datos como Int2DoubleOpenHashMap.

Sobre RankSys se han implementado distintas variantes del algoritmo knn:

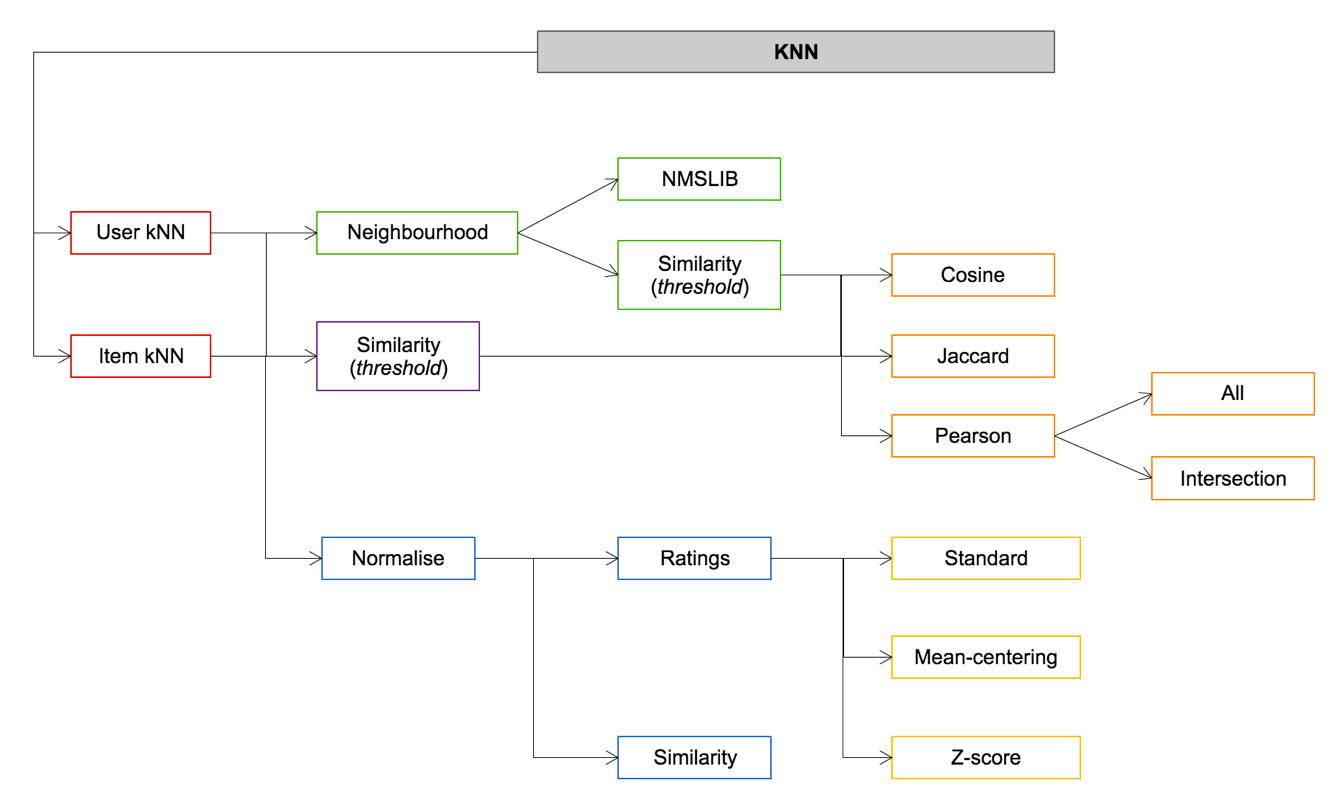


Figura . Esquema de las variantes en kNN

Si tenemos en cuenta la fórmula del cálculo de ratings por ejemplo en filtrado colaborativo, podemos realizar tres modificaciones generales tal como se muestra en el diagrama:



En cuanto al vecindario, puede ser obtenido por dos vías, recibiendo los vecinos de la librería NMSLIB, explicado en [3.4.1](#_Librerías_externas_utilizadas) y [3.4.2](#_Cálculo_aproximado_de) o bien por el método tradicional calculando la similitud entre los usuarios y obteniendo las k similitudes más altas.

Dentro de las similitudes existen diferentes opciones como se menciona en 2.1.2. Dos de ellas, coseno y Jaccard son utilizadas directamente desde la implementación que ofrece RankSys, mientras que Pearson es implementada desde el inicio ofreciendo a su vez otras dos variantes. También existe la posibilidad de aplicar un *threshold* a la similitud, esto es un valor límite por debajo del cual no se van a considerar las similitudes, en el caso de escoger un threshold de 0.3, el vecindario no contendrá vecinos con un valor de similitud por debajo de 0.3, únicamente valores superiores.

Cuando se calcula una similitud entre dos usuarios se opera sobre los ítems que ambos han valorado, aunque también cabe la posibilidad de utilizar todo el conjunto de ítems, asignando un valor por defecto a los ítems no comunes. Normalmente este valor suele ser 0, pero un valor interesante podría ser la media del usuario para evitar distorsionar en exceso la similitud.

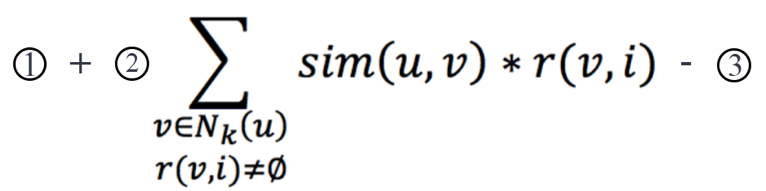
En el diagrama, “Pearson all” hace referencia a esto último, mientras que Pearson intersection implica el conjunto de ítems puntuados en común.

La segunda variante (Similarity threshold) puede parecer redundante, ya que se ha utilizado antes para el vecindario, pero no es así, pues aunque normalmente utilicemos la misma similitud para calcular los vecinos y la que multiplica al rating, según la fórmula *sim*(*u,v*), es posible utilizar diferentes para cada caso, habiendo las tres posibilidades anteriores con la posibilidad de aplicar threshold.

La tercera y última variación es la normalización, bien de ratings ([2.2](#_Normalización_de_ratings)) o de similitud. La normalización de similitudes implica dividir por el valor absoluto de la suma de las similitudes utilizadas en el sumatorio. Por otra parte, la normalización del rating implica establecer un intervalo fijo en el que se van a representar, normalmente [1,5]. En el caso de no aplicar normalización al rating, los valores de recomendación en vez de variar dentro del rango, tomarían valores muy altos, resultante de la suma de todos los ítems comunes.

Lo más común es que la normalización se aplique únicamente para presentarle al usuario los resultados, ya que internamente, no importa tener valores desproporcionados ya que hay que ordenarlos, siendo el valor más alto el ítem más recomendable.

De esta manera podríamos parametrizar la formula en tres puntos:



* El primero. En el primero se suma la media del usuario en Mean-Centering, mientras que, en el segundo, usado en Z-score ([2.2](#_Normalización_de_ratings_1)) se calcula la suma de la media y la desviación típica del usuario.
* En el segundo, se decide normalizar o no la similitud.
* En el último se aplica la normalización de ratings, restando 0 en caso de STD, la media del usuario en Mean-Centering y la media del usuario dividida entre la desviación típica si aplicamos Z-score.

### Librerías externas utilizadas

La librería principal en la que está basado el proyecto es RankSys2, un framework para la implementación de algoritmos de recomendación y evaluación. RankSys ha sido programada en Java 8, tomando ventajas de ésta nueva versión, como las expresiones lambda o *streams*, así como la facilidad de paralelizar el código para obtener mayor rendimiento. Se han evaluado otras librerías como LensKit o Mahout pero no se obtenían los mismos resultados que con RankSys.

RankSys ofrece una base a la de recomendación y evaluación, posee

algoritmos báscos (popularidad, knn, factorización de matrices), similitudes (coseno, jaccard) para métodos basados en ítem y usuario.

La otra librería utilizada se trata de NMSLIB, la cual realiza un cálculo de vecinos de manera aproximada para ahorrar tiempo de ejecución y recursos hardware, pues en lugar de ordenar la lista de todos los vecinos para saber los k más cercanos, calcula la distancia entre los vectores (usuarios o ítems) en un espacio multidimensional.

### Cálculo aproximado de vecinos

El cálculo de vecinos aproximados nace con la librería Annoy3 creada por Erik Bernhardsson [2], [3] y [4]. La finalidad es la misma que el cómputo mediante similitudes, obtener los k vecinos más cercanos a un usuario en concreto.

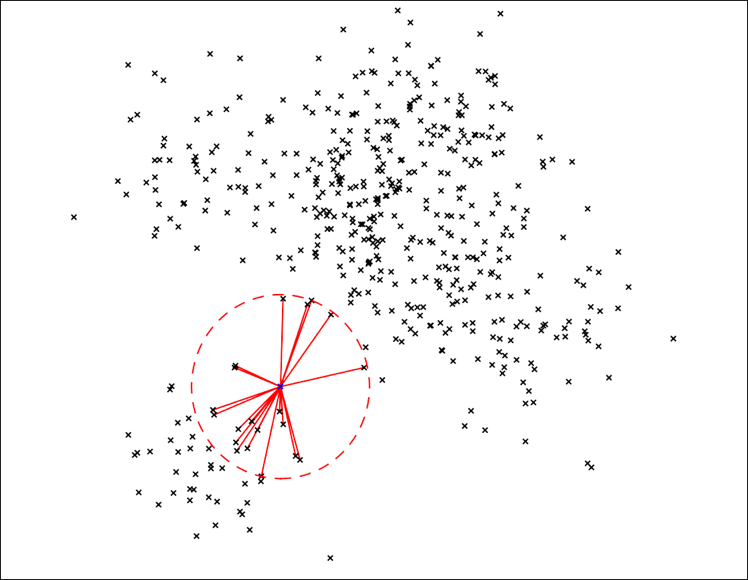
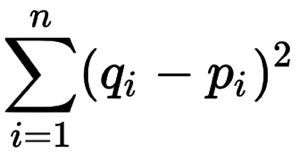


Figura . Mapa bidimensional de usuarios

La idea básica es representar objetos en un espacio donde la proximidad implique que sean parecidos entre ellos. En el caso de filtrado colaborativo, un dataset con 10 usuarios y 5 ítems compondrá un espacio de 50 dimensiones. La proximidad entre dos vectores normalmente se mide con la distancia euclídea al cuadrado:



Por lo que en éste método se excluyen las similitudes y tomamos como valor de proximidad la distancia entre los vectores.

Volviendo al mapa de usuarios, el objetivo es calcular los vecinos más cercanos en el espacio. Nótese que se representa un espacio de dos dimensiones por un motivo meramente visual, ya que en realidad la mayoría de modelos vectoriales aplicados a sistemas de recomendación poseen dimensiones mucho más grandes.

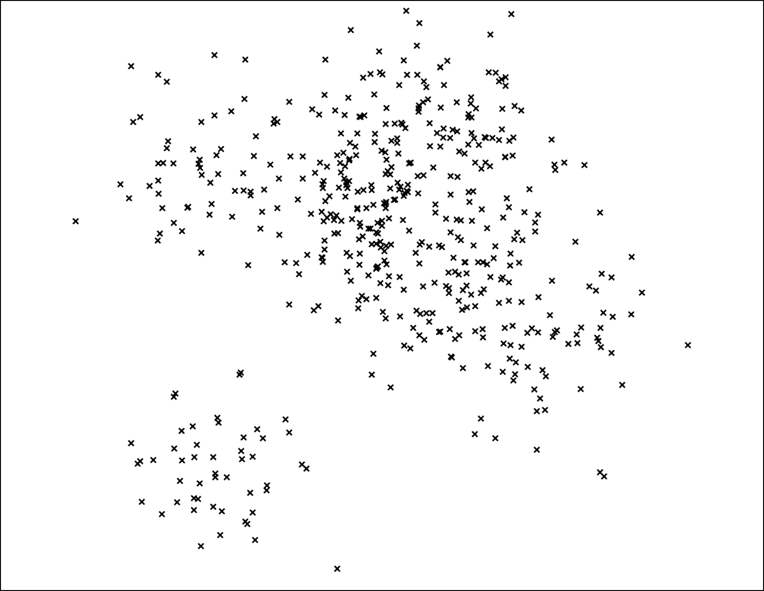


Figura . Mapa de distribución de usuarios en el espacio

El primer reto es establecer una estructura de datos que permita encontrar las distancias más cortas a un vector en un tiempo sublineal. Parece natural elegir un árbol para ello, pues su rendimiento es de orden *O*(*log n*).

Ésta es la manera en que Annoy funciona, un árbol binario donde cada nodo es una partición aleatoria del conjunto de usuarios, de manera recursiva, se van tomando dos puntos (usuarios) de manera aleatoria y se traza un hiperplano equidistante a ambos.

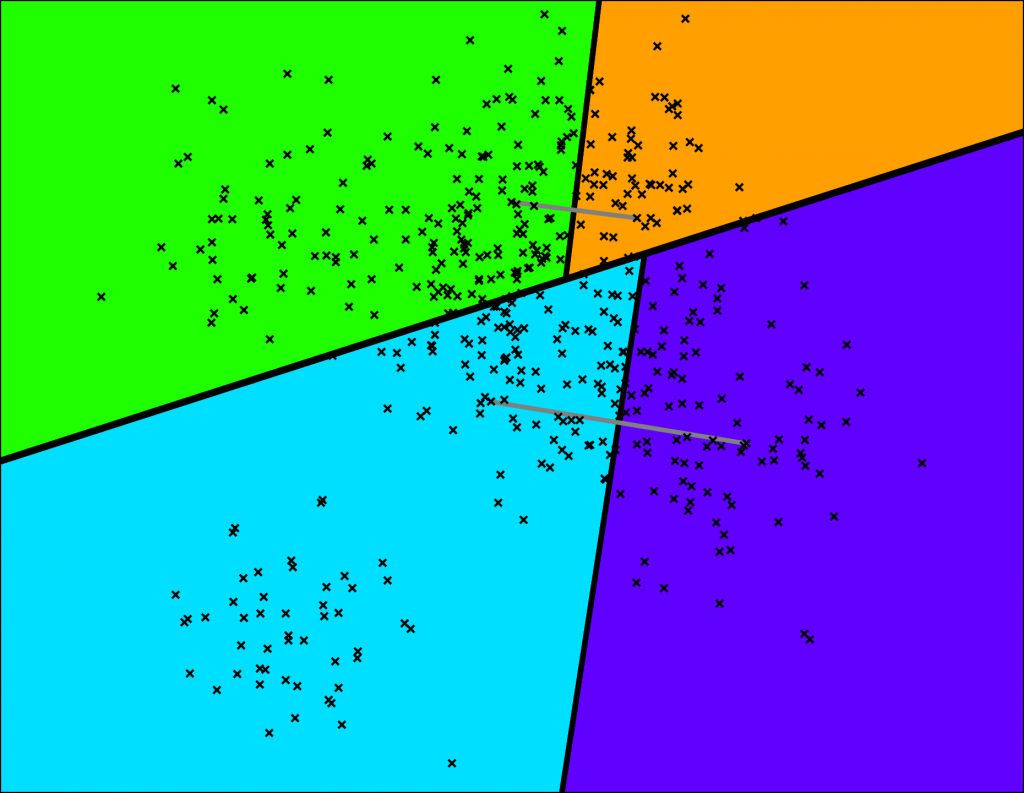


Figura . Mapa con tres particiones aleatorias

En este punto el árbol resultante sería el siguiente:

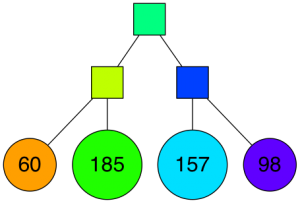


Figura . Árbol resultante de realizar tres particiones

Tras realizar sucesivas iteraciones hasta que en cada nodo queden 10 usuarios como mucho, nos quedan las siguientes particiones:

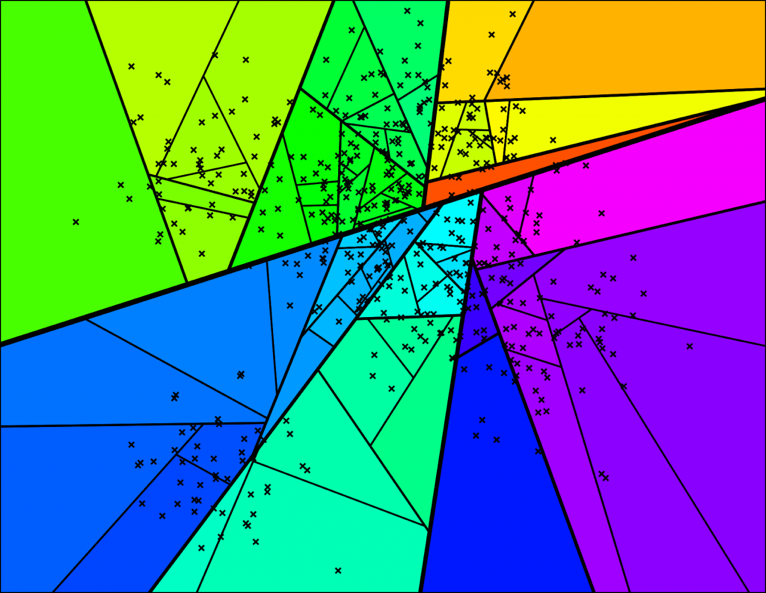


Figura . Mapa de particiones para k=10

Y el correspondiente árbol:

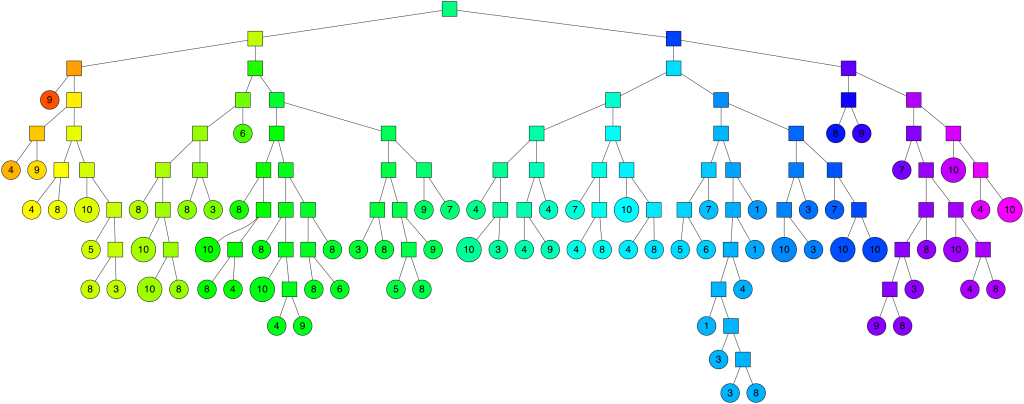


Figura . Árbol binario de particiones para k=10.

Una vez completado el árbol, se puede observar que, si dos puntos están cerca en el espacio, también estarán cerca en el árbol. Para buscar cualquier punto en el espacio, simplemente se ha de recorrer el árbol desde la raíz. Cada punto intermedio definidos como cuadrados, representan un hiperplano, de manera podemos saber hacia qué hijo deberemos dirigirnos a la hora de buscar.

Imaginemos que buscamos un usuario concreto que resultar ser el siguiente:

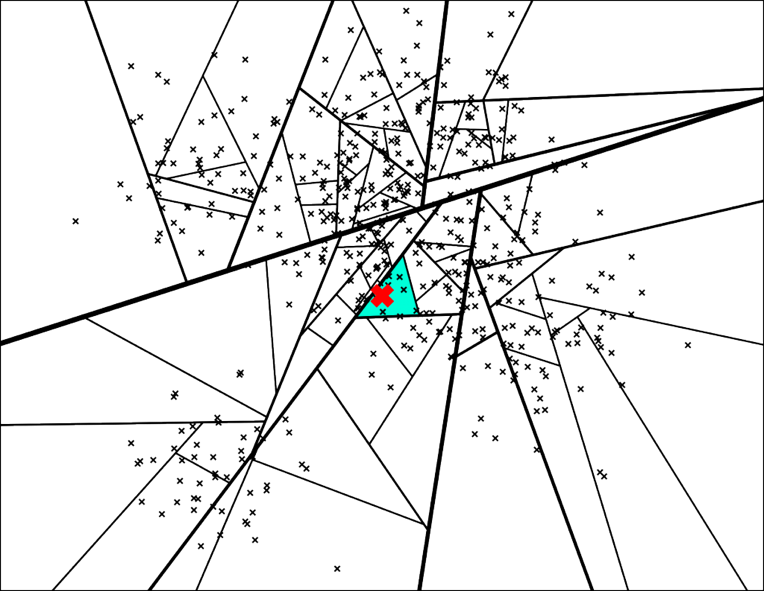


Figura . Búsqueda de un usuario en el espacio

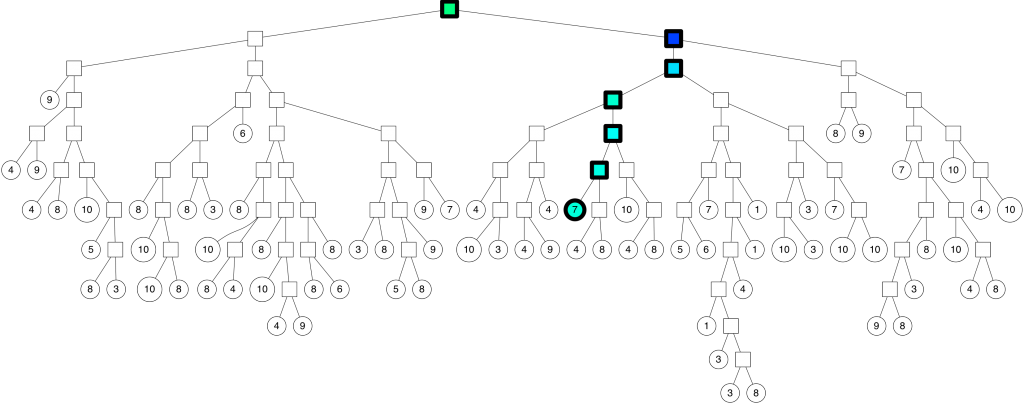


Figura . Árbol resultante de la búsqueda de un usuario

En el área que contiene al usuario en el espacio hay siete vecinos. Si observamos, éstos no resultan ser los más cercanos, pues el usuario está en la zona delimitada por un hiperplano quedando vecinos más cercanos en un área distinta.

Por lo tanto, hay dos problemas a resolver:

* Encontrar los vecinos realmente más cercanos
* Encontrar más de los siete vecinos que definen el área encontrada.

La primera solución pasa por ir explorando las áreas colindantes, profundizando en el árbol estableciendo un threshold (umbral) en las distancias para saber si estamos en la rama correcta. Para ello se utiliza una cola de prioridad para explorar nodos ordenados por la distancia al objetivo.

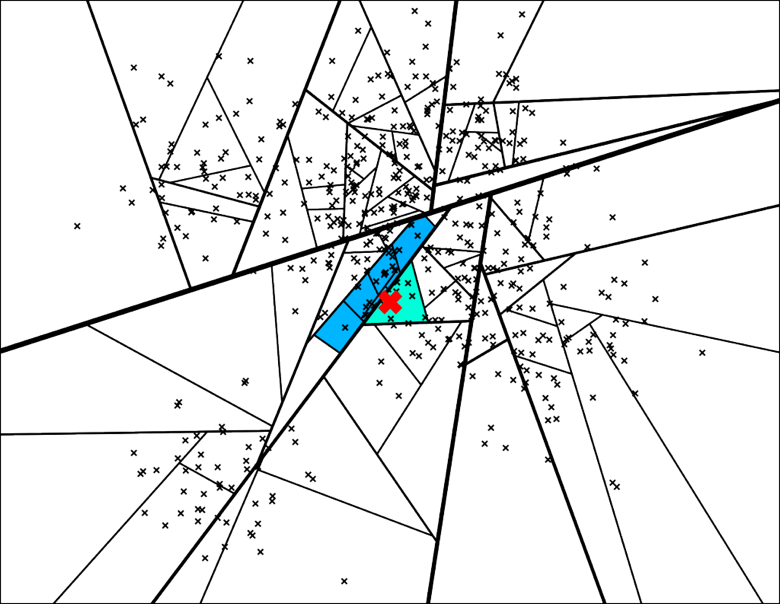


Figura . Mapa de ampliación en la búsqueda de vecinos

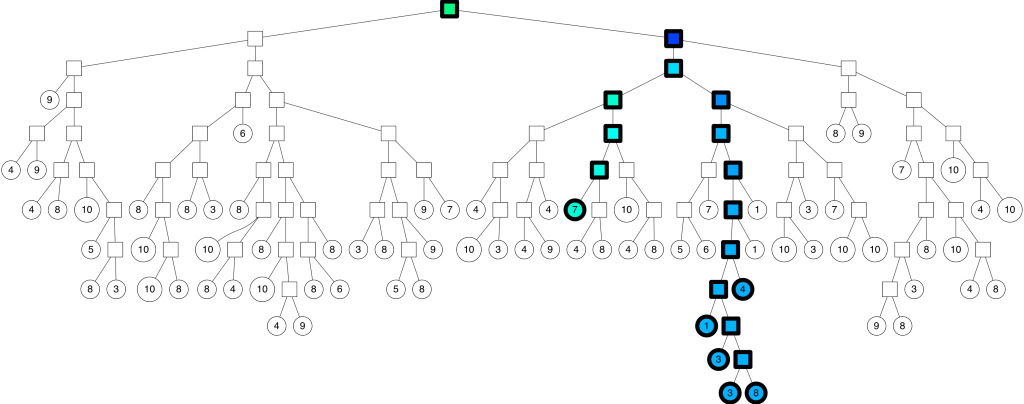


Figura . Árbol de ampliación en la búsqueda de vecinos

Otra posible solución pasa por crear muchos árboles (bosque), cada uno construido con distintas particiones, los cuales van a ser recorridos en la búsqueda al mismo tiempo.

Como tenemos la cola de prioridad ya configurada, es posible recorrer todos los árboles para obtener la unión de las hojas de los árboles (áreas) que contienen los puntos más cercanos. Ésta unión de áreas representa el vecindario:

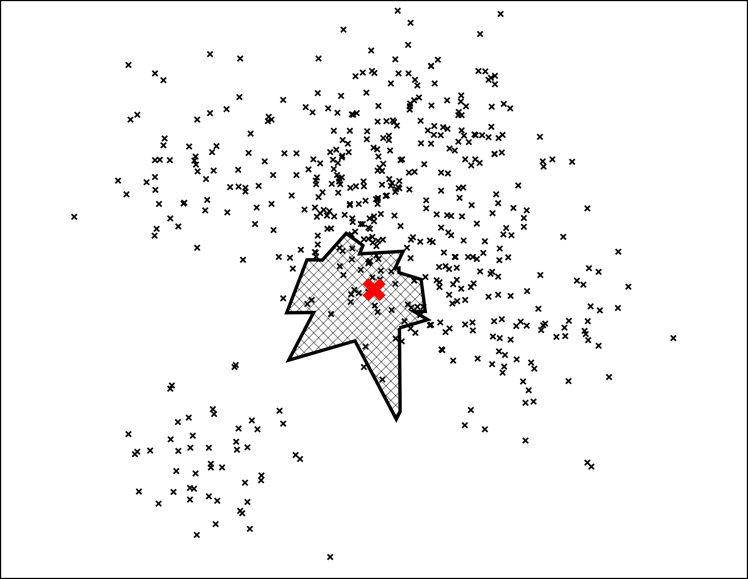


Figura . Vecindario de candidatos a vecinos más próximos en el espacio

Una vez obtenido el conjunto de candidatos, hay que calcular las distancias entre ellos, ordenándolas para obtener los más cercanos. Aunque si observamos con detenimiento, hay vecinos que se han escapado del área:

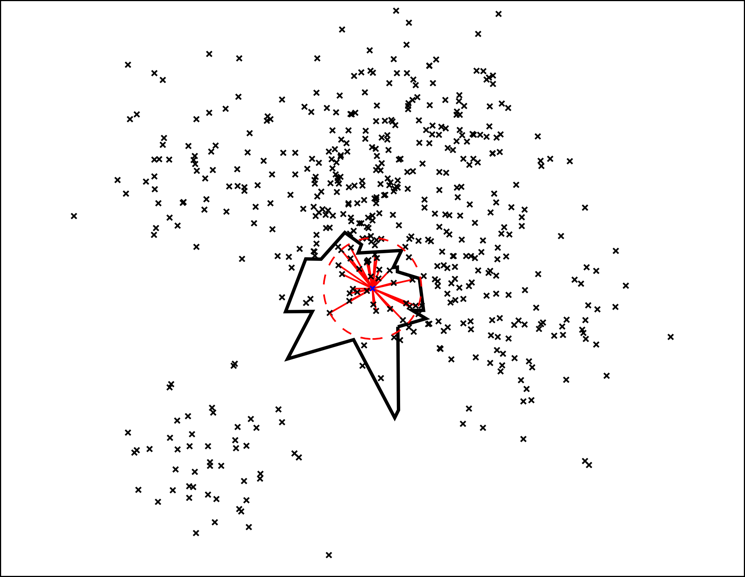


Figura . Radio que comprende los vecinos más próximos en el espacio

Por ello se denomina cálculo *aproximado* de vecinos, siempre existe la probabilidad de que queden excluidos vecinos realmente relevantes. La filosofía de este método yace en que, sacrificando un poco de precisión, somos capaces de obtener una gran mejora (en torno a 100 veces mejor) en el rendimiento cuanto más grande sea el conjunto de usuarios o ítems. Existe la posibilidad de aumentar la precisión, esto es, aumentando el número de árboles, pero perderíamos eficiencia en cuanto a memoria utilizada, por lo que hay que encontrar el compromiso entre estos factores.

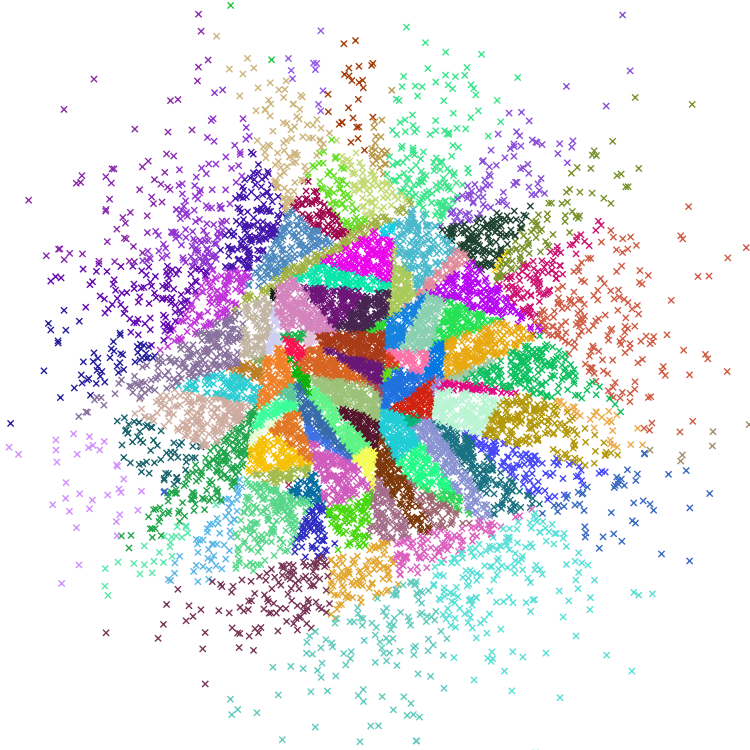


Figura . Mapa bidimensional de vecinos con particiones aleatorias

### Annoy vs NMSLIB

La razón por la que en este proyecto se incluye la librería NMSLIB en vez de Annoy es debido a que, en mayo de 2016, NMSLIB le ha superado notablemente en términos de rendimiento y precisión.

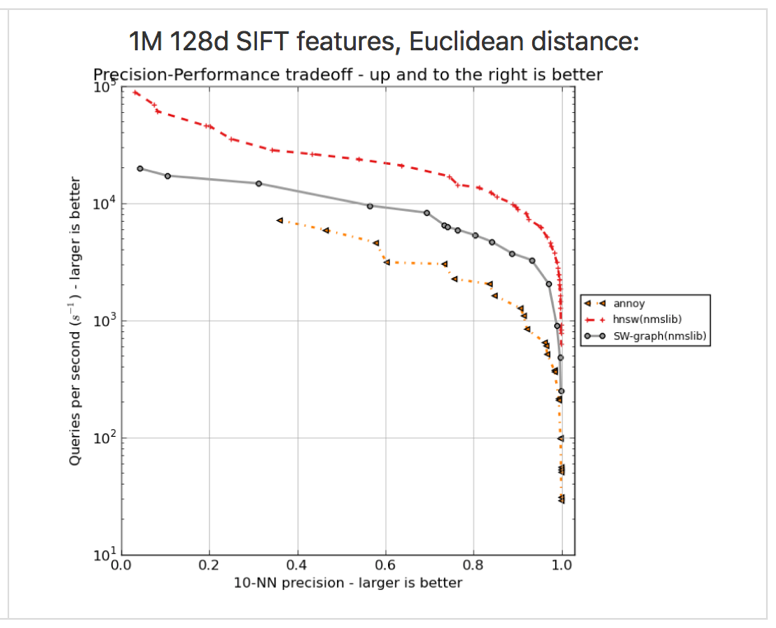


Figura . Annoy vs NMSLIB

### Apache Thrift y NMSLIB

El principal problema del uso de la librería NMSLIB es que está escrita en el lenguaje C++, mientras que el proyecto es Java, lo que requiere de algún método para compatibilizar ambos lenguajes.

Mediante Apache Thrift, un framework que permite la creación y definición de servicios en múltiples lenguajes, usando RPC (*Remote Procedure Call*) y generación de código, es posible conectar aplicaciones escritas en multitud de lenguajes diferentes.

De esta manera, se instancia un servidor escrito en C++ con la librería NMSLIB y un cliente en Java que se conectará al servidor para ejecutar los métodos necesarios de manera remota (RPC). En este caso el servidor realizará el cálculo de vecinos devolviendo una lista de usuarios o ítems al cliente de manera totalmente transparente gracias a Thrift.

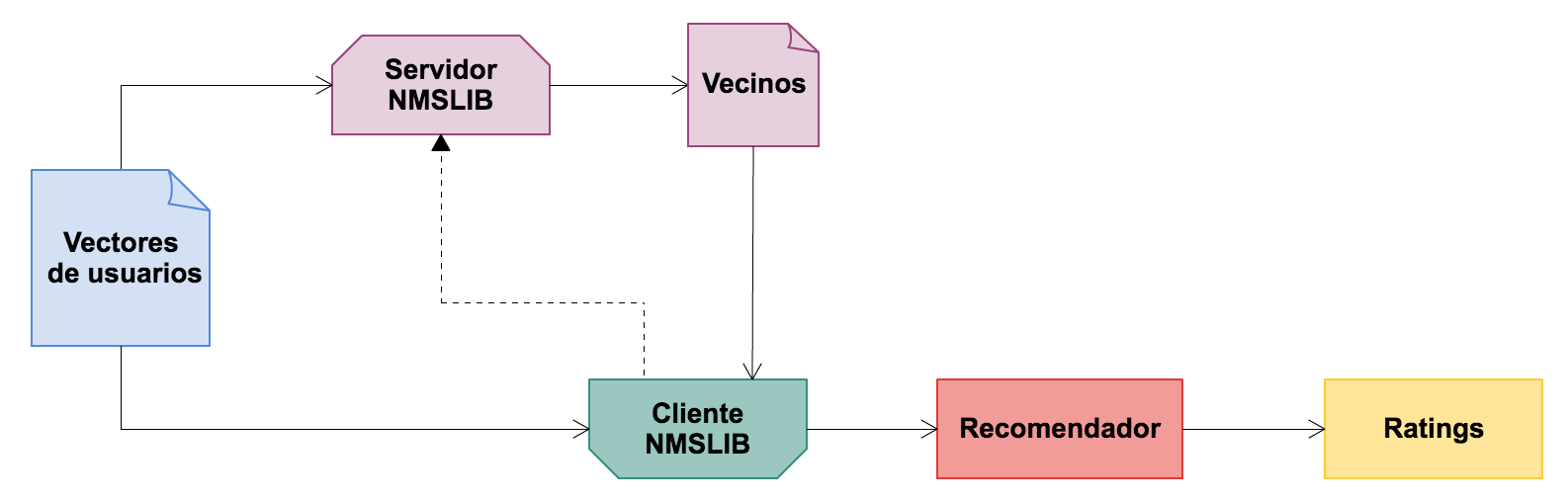


Figura . Esquema de empleo de NMSLIB

Una vez obtenidos los vecinos de por NMSLIB, son tratados por el programa cliente exportándolos a un fichero, el cuál será la entrada del recomendador para la generación de recomendaciones.

# Integración, pruebas y resultados

# Conclusiones y trabajo futuro

….

Ya ha quedado claro que los sistemas de recomendación son realmente útiles, tanto para usuarios como comercios online. Un recomendador ofrece un gran abanico de posibiliades, desde incrementar el número de ventas o clientes, hasta proporcionar una mejor experiencia de usuario, hasta el punto de no preocuparse de buscar entre millones de artículos, teniendo al alcance recomendaciones totalmente personalizadas.

Durante este trabajo se han ido desarrollando las facetas más básicas de los sistemas de recomendación, profundizando a su vez en el algoritmo de vecinos cercanos

# Referencias

1. A. Bellogín, Recommender System performance evaluation and prediction: An information retrieval perspective, Thesis, October 2012, pp 17-35.
2. E. Bernhardsson, Algorithms and data structures, September 2015.
3. E. Bernhardsson, Curse of dimensionality, September 2015.
4. E. Bernhardsson, Nearest neighbors and vector models, September 2015.
5. F. Ricci, L. Rokach, B. Shapira, Paul B. Kantor (eds.). Recommender Systems Handbook, 1st edition, Springer, 2011, Chapter 1. Introduction to Recommender Systems Handbook.
6. F. Ricci, L. Rokach, B. Shapira, Paul B. Kantor (eds.). Recommender Systems Handbook, 1st edition, Springer, 2011, Chapter 3. Content-based Recommender Systems: State of the Art and Trends.
7. F. Ricci, L. Rokach, B. Shapira, Paul B. Kantor (eds.). Recommender Systems Handbook, 1st edition, Springer, 2011, Chapter 4. A Comprehensive Survey of Neighborhood-based Recommendation Methods.
8. F. Ricci, L. Rokach, B. Shapira, Paul B. Kantor (eds.). Recommender Systems Handbook, 1st edition, Springer, 2011, Chapter 8. Evaluating Recommendation Systems.
9. P. Sanchez, Estudio y aplicación de algoritmos y estructuras de datos a los Sistemas de Recomendación, Trabajo Fin de Grado, Escuela Politécnica Superior, Universidad Autónoma de Madrid, May 2016.
10. Premio netflix. <http://www.netflixprize.com/>
11. S. Marina, Suite de algoritmos de recomendación en aplicaciones reales, Trabajo Fin de Grado, Escuela Politécnica Superior, Universidad Autónoma de Madrid, May 2014.

# 

# Glosario

* **Annoy**: Approximate nearest neighbours. Librería que calcula vecinos de manera aproximada.
* **Apache Thrift**: Framework
* **Dataset**:Conjunto de datos analizados
* **DCG**: Discounted cumulative gain. Métrica de evaluación.
* **Framework**:
* **iDCG**: ideal DCG. Métrica de evaluación.
* **kNN**: k nearest neighbours. k vecinos más cercanos.
* **MAE**: Mean absolute error. Métrica de evaluación.
* NDCG: Normalised discounted cumulative gain
* **NMSLB**: Non-metric space library. Librería basada en Annoy para el cálculo aproximado de vecinos.
* **RankSys**: Ranking system. Librería base del proyecto, aporta funcionalidad de generación de recomendaciones y evaluación.
* **Rating**: Puntuación que un usuario da a un ítem.
* **Recall**: Métrica de evaluación.
* **RMSE**: Root mean squared error. Métrica de evaluación.
* **RPC**: Remote procedural call. Llamada a procedimiento remoto, usado para ejecutar código en otra máquina (Thrift)
* **Split**: Partición del dataset, dividiendo en training y test.
* **SR**: Sistema de recomendación.
* **Test**: Partición que contiene las recomendaciones a evaluar.
* **Tf-idf**: Term frequency-inverse document frequency. Métrica de frecuencia de términos.
* **Threshold**: Umbral por debajo del cual las similitudes son descartadas.
* **Training**: Partición del dataset utilizada para “enseñar” al sistema de recomendación.

# Anexos

## Manual de instalación

Instalación de apache thrift y librerías para usar nmslib

1. Desde el repositorio en GitHub de NMSLIB1 se descarga el .zip del proyecto

Dependiendo del repositorio disponible instalar una de las dos:

sudo apt-get install libboost1.58-all-dev

sudo apt-get install libboost1.54-all-dev

Instalar las siguientes librerías:

sudo apt-get install libgsl0-dev libeigen3-dev

sudo apt-get install cmake

sudo apt-get install libboost-dev libboost-test-dev libboost-program-options-dev libboost-system-dev libboost-filesystem-dev libevent-dev automake libtool flex bison pkg-config g++ libssl-dev libboost-thread-dev make

En la carpeta similarity\_search ejecutar: cmake . ; make

1. Descargar Apache Thrift2 (actualmente la librería funciona con Thrift 0.9.2).

Para instalar:

./configure

sudo make

sudo make install

Verificar instalación: thrift -version

1. Una vez están instalados todos los paquetes necesarios se puede compilar el servidor (directorio query\_server/cpp\_client\_server): make
2. Y ejecutar el servidor (p.ej):

./query\_server –i inputFile -s l2 -m sw-graph -c NN=10,efConstruction=200,initIndexAttempts=1 -p 10000

NOTA: si al ejecutar el servidor salta el error “./query\_server: error while loading shared libraries: libthrift-0.9.2.so: cannot open shared object file: No such file or directory”. Ejecutar sudo ldconfig

Antes de ejecutar el cliente es necesario compilar. Ya que es un proyecto maven, necesitaremos descargar las dependencias necesarias (especificadas en el pom.xml) para ello ejecutamos: mvn package

Antes de ejecutar el cliente necesitamos un fichero de datos:

export DATA\_FILE=../../sample\_data/final8\_10K.txt

Para ejecutar el cliente:

exec:java -Dexec.mainClass=thrift.Client -Dexec.args="-p 10000 -a localhost -k 500 -i inputFile(usersRatingsVector) -out utFile(neighbours)"

Nos devolvería los 500 vecinos más cercanos al primer usuario

NOTA: tener inicializada la variable JAVA\_HOME para ejecutar el cliente.

## Manual del programador

## Anexo …

## 