ESCUELA SUPERIOR POLITECNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

"Sistema de predicción y recomendación personalizada basada en ranking de ítems homogéneos usando filtrado colaborativo"

TESIS DE GRADO

Previo a la Obtención del Título de:

INGENIERO EN COMPUTACIÓN
ESPECIALIZACIÓN EN SISTEMAS TECNOLÓGICOS
INGENIERO EN COMPUTACIÓN
ESPECIALIZACIÓN EN SISTEMAS DE INFORMACIÓN
INGENIERO EN COMPUTACIÓN
ESPECIALIZACIÓN EN SISTEMAS DE INFORMACIÓN

Presentada por:

Chang Miranda Hugo Iván
Díaz Viejó Luís Alejandro
Ruiz Moncayo Fausto Daniel

GUAYAQUIL – ECUADOR
AÑO - 2006

AGRADECIMIENTO

A todas las personas que de uno u otro modo colaboraron en la realización de este trabajo y especialmente al Msc. Fabricio Echeverría, Director del Tópico, Ing. Verónica Macías y al Ing. Carlos Jordán, vocales.

DEDICATORIA

A Dios y mi querida Mater, por los talentos y capacidades que me han regalado y que me han permitido cristalizar este proyecto de tesis.

A mi familia, mis padres y hermanos, por ser mi ejemplo, ser mi fortaleza y la alegría de mis días.

A Isabel, por caminar conmigo aún en los momentos difíciles y mostrarme la fuerza que tiene tu amor y tu abrazo.

A mis amigos, por cada experiencia compartida en este camino de crecer y llenar mi vida de hermosos recuerdos.

Para ustedes mi entrega y gratitud.

Fausto Ruiz

A Dios por permitirme culminar satisfactoriamente este proyecto de tesis y haberme dado la salud y constancia para terminar mi carrera y por las bendiciones que me da.

A mis padres; a mi madre, María Miranda Medina por su apoyo incondicional, su paciencia y su fe; a mi padre, Víctor Hugo Chang, que a pesar de nuestras diferencias, estas se han encauzado a templar mi carácter en la consecución de mis metas y jamás darme por vencido.

A mis hermanos, Kenya y Douglas; a mi querida tía Blanca por todo su cariño y apoyo; a mi adorada Katiuskita por su paciencia y amor y a todos quienes a lo largo de la carrera han aportado directa e indirectamente en mi formación: profesores, compañeros y amigos (Patricio, Paulo y Alfredo); y un especial agradecimiento a mis compañeros de proyecto, Alejandro Díaz y Fausto Ruiz, por su apoyo, trabajo y consideración.

"La meta no es lograr ser perfectos sino lograr ser menos imperfectos"

Hugo Chang

A mi hermana Gina Díaz de Dueñas y al Dr. Marlon Dueñas, su apoyo fue imprescindible para la culminación de mi carrera.

A Mónica Pibaque su amor incondicional me lleva hacia adelante.

A mis hermanos Marlon y Andrés, mis mejores amigos.

Luís Díaz

TRIBUNAL DE GRADUACIÓN

Ing. Holger Cevallos SUBDECANO DE LA FIEC PRESIDENTE	Msc. Fabricio Echeverría DIRECTOR DE TÓPICO
Ing. Verónica Macías MIEMBRO DEL TRIBUNAL	Ing. Carlos Jordán MIEMBRO DEL TRIBUNAL

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de esta Tesis de Grado, nos corresponden exclusivamente; y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL"

(Reglamento de Graduación de la ESPOL)

Fausto Daniel Ruiz Moncayo

Luis Alejandro Díaz Viejó

Hugo Iván Chang Miranda

RESUMEN

Actualmente, vivimos en un medio donde la oferta y demanda de productos y servicios se mueve a grandes velocidades, vivimos en un mercado en constante movimiento que ofrece al consumidor nuevas opciones de compra cada día. Este panorama representa un reto para las personas y/o empresas que ofertan sus productos, al tener que satisfacer las necesidades, cada vez más personalizadas, de los consumidores.

El consumidor se ve abrumado por el abanico de posibilidades que se le presenta al momento de elegir un determinado objeto cultural, entiéndase por ello a objetos de uso cotidiano para el esparcimiento, alimentación, entretenimiento, entre otros. Un problema evidente surge a la vista: cada vez será más difícil escoger un objeto cultural como una película, un libro o un restaurante, que sea realmente de nuestro interés y que, además de ir acorde a nuestro gusto, satisfaga nuestra expectativa de calidad. Usando una analogía, que emplearemos durante del desarrollo del proyecto de tesis, diríamos que es como “encontrar una aguja en un pajar”.

Consideramos que es posible alcanzar un beneficio importante en la tarea de elección de aquello que realmente busco y que será de mi agrado, por

medio del uso de la tecnología y las técnicas de la minería de datos. Con esta motivación presentamos este proyecto de tesis.

Los sistemas de predicción y recomendación forman parte de un grupo de nuevas tecnologías que se encuentran dentro del estudio de la minería de datos, la cual intenta aprovechar el conocimiento que poseen ciertas personas acerca de determinados objetos culturales para ayudarse entre sí, generar nuevo conocimiento y darles a conocer nuevos objetos que posiblemente sean de su interés.

Los sistemas de predicción, hacen uso de mecanismos naturales de búsqueda que seguro hemos utilizado; en más de una ocasión, hemos tomado decisiones sobre la elección de un producto o servicio basado en la opinión de una persona o grupo de personas en quienes confiamos, de manera que objetos recomendados por otros (un libro, una película, una obra musical) nos estimulan a conocerlos.

La meta que perseguimos con el presente proyecto es presentar las características y funcionamiento de los algoritmos de predicción bajo el uso de filtrado colaborativo a través del desarrollo de “AFI Restaurantes”, un sitio Web que represente un sistema de recomendación.

Los usuarios registrados en el sistema votarán o calificarán los objetos culturales que conozcan; esta información será utilizada por el sistema para predecir y agrupar usuarios con comportamientos y gusto similares; y se recomendará a cada individuo aquellos objetos que tuvieron una valoración positiva en los usuarios más parejos a él. Para el desarrollo del proyecto se escogió como ítem cultural los restaurantes de Guayaquil.

En cada uno de los capítulos de la tesis iremos analizando y desarrollando el contenido que soporta nuestro proyecto.

En el primer capítulo, se expone el planteamiento formal del problema, objetivos del proyecto y se estudian ventajas y desventajas del uso del filtrado colaborativo.

En el segundo capítulo, se introduce un fundamento teórico acerca de los sistemas de recomendación y predicción.

En los capítulos tres y cuatro, se describe detalladamente el análisis y diseño del proyecto, incluye especificaciones UML y la justificación de los algoritmos escogidos.

En el capítulo cinco, se hace una descripción de la metodología de implementación del proyecto. Dentro de este capítulo se exhiben las técnicas de análisis multivariante que usaremos y se enfatiza sobre la implantación e impacto de la minería de datos.

En el capítulo seis, se realiza un análisis financiero de la solución, los costos, las oportunidades y los productos similares existentes en el mercado. Y finalmente, se expondrán las conclusiones y recomendación de la presente tesis de grado.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

Planteamiento del problema

1.1. Definición del problema

Nos encontramos en la actualidad viviendo en una era donde la oferta de ciertos objetos culturales es cada vez más dinámica y extensa, dando origen a una brecha entre la cantidad y calidad que el consumidor final percibe. En numerosas ocasiones, habremos encontrado el problema de no poder elegir una película que sea de nuestro interés, un libro de tantos que se publican anualmente o también el poder elegir un buen restaurante.

En la práctica y en nuestra vida cotidiana, la selección de un producto o servicio que puede ser de nuestro interés se vuelve una tarea muy complicada debido a la inmensa variedad de opciones que se genera a nuestro alrededor. Haciendo una analogía, diríamos que es como “encontrar una aguja en un pajar”.

1.2. Objetivos del proyecto

Teniendo claramente definido el problema nos sugerimos la siguiente interrogante: ¿Puede la tecnología ayudar de alguna manera a las personas a elegir realmente lo que buscan? ¿Ayudar a "encontrar la aguja en el pajar"? Pensamos que sí, y esto ha marcado el origen de este proyecto de tesis. Durante el desarrollo del mismo perseguiremos los siguientes objetivos:

- Exponer las características y funcionamiento de los algoritmos de predicción basados en ítems y, tomando como fundamento las técnicas de la minería de datos, poder mostrar su aplicación específica en la tarea de predicción y recomendación.
- Desarrollar e implementar un sitio Web que se base en el uso de algoritmos de filtrado colaborativo y solucione el problema de hacer recomendaciones personalizadas de ítems de un mismo tipo (homogéneos) a usuarios registrados que presentan comportamientos y gustos similares. El dominio de producto o servicio escogido es el de los restaurantes de Guayaquil.
- Definir las ventajas y retos del uso del filtrado colaborativo en la elaboración de recomendaciones, así como también métodos y sugerencias para garantizar la escalabilidad en ambientes de producción que manejan grandes volúmenes de datos.

1.3. Análisis contextual y dominio del producto

1.3.1. Objetivos del análisis

Como primer paso en el desarrollo del presente proyecto plantearemos un análisis del ámbito donde se desarrolla nuestro sistema y nuestros usuarios. Respecto al entorno analizaremos la realidad económica y social de la ciudad con el objetivo de comprender el ritmo de la actividad de restaurantes en la urbe. Respecto a los usuarios nos interesa estudiar los cambios de su comportamiento a través de los años para establecer cuáles son los factores que son determinantes al momento de tomar decisiones.

1.3.2. Análisis del usuario y su entorno

Para comprender a nuestros usuarios debemos analizar qué es lo que los motiva al momento de adquirir un bien o servicio. Como punto de partida se tomará a consideración la tendencia de consumo de los ecuatorianos.

De acuerdo a la última encuesta realizada por el Instituto Ecuatoriano de Estadística y Censos (INEC), hace 10 años los ecuatorianos gastaban el 32% de sus ingresos en alimentos y

bebidas no alcohólicas[1], entendemos con esto que nos estamos refiriendo a bienes de primera necesidad. En el 2005, esta proporción se redujo al 27.3%, ver Figura 1 - Distribución de la canasta de consumo nacional (2006)

Figura 1 - Distribución de la canasta de consumo nacional (2006)[2]

[1] INEC – Encuesta anual de hoteles, restaurantes y servicios.
[2] "Los gastos suntuarios de los ecuatorianos en crecimiento", Diario Expreso, 23 de mayo del 2006.

Estas cifras nos indican que el comportamiento del consumidor a la hora de elegir los productos que satisfagan sus necesidades ha cambiado; y además, cada vez los individuos destinan más a los gastos suntuarios como: la telefonía celular, el alcohol, cigarrillos, televisión pagada, cine y restaurantes de comida rápida. En paseos, entretenimiento y diversión, sumado a los desembolsos en restaurantes y comidas rápidas, el ecuatoriano gasta aproximadamente el 8% de sus ingresos.[3]

Ante esta tendencia, los estudios de mercado para los establecimientos de comida se ven muy alentadores al indicar una clara predisposición: cultural, social y económica; para el consumo de sus bienes y servicios.

El sociólogo Roberto Sáenz, director regional del litoral del INEC, anota que "los patrones del comportamiento del consumidor han cambiado significativamente, influenciadas por el exceso de publicidad." Según expresa, hay una teoría de marketing que dice, "crea la necesidad y crea el estilo de vida."

"La gran masa de publicidad ha hecho que las personas adquieran otro patrón de consumo y sientan la necesidad de

[3] "Los gastos suntuarios de los ecuatorianos en crecimiento", Diario Expreso, 23 de mayo del 2006

estar a la “onda”, consumiendo en productos que no son necesariamente los prioritarios”, enfatiza el experto.[3]

Socialmente, en la actualidad se puede apreciar que han desaparecido –en parte- las jerarquías culturales. El usuario no tiene referentes, ya que las recomendaciones “críticas” están a su vez sobrecargadas y/o manipuladas por intereses comerciales.

Siendo ésta la realidad, podemos afirmar que una herramienta que ayude a tomar decisiones de manera más objetiva, beneficiará al consumidor, quien espera que el producto adquirido cumpla con sus expectativas iniciales. Al satisfacer esta necesidad, se estimulan compras futuras, se mueve mercado y se alcanza fidelidad a productos y servicios; logrando así, un escenario de beneficio mutuo entre consumidor y proveedor.

1.4. Ventajas del filtrado colaborativo

- Los sistemas de recomendación ha probado ser una importante respuesta al problema de la sobrecarga de información que sufre el usuario de objetos culturales al proveer de servicios dinámicos y de información personalizada.

- Las técnicas de filtrado colaborativo [4] han probado ser un componente vital de muchos sistemas de recomendación al facilitar la generación de recomendaciones de alta calidad al poder predecir las preferencias de comunidades de usuarios similares.

Encontraremos, además de estas ventajas generales, algunas ventajas específicas de cada enfoque de implementación del filtrado colaborativo que vienen dadas por las bondades que nos ofrecen los diferentes algoritmos existentes como el filtrado colaborativo basado en: la memoria, el basado en usuarios y el basado en ítems.

1.5. Desventajas del filtrado colaborativo

Los sistemas de recomendación basados en usuarios han sido muy exitosos en el pasado, pero su extensivo uso actual ha revelado potenciales retos y/o desventajas que deberemos tener en cuenta al desarrollar nuestro sistema:

- **Esparcidad:** Se refiere a la medida de dispersión de los datos que serán usados para el análisis y procedimientos de predicción. En grandes sistemas de recomendación tales como Amazon.com (recomienda libros) y CDnow.com (recomienda álbumes de música), en donde se manejan grandes volúmenes de

[4] **Filtrado Colaborativo:** algoritmo usado por técnicas de la minería de datos para analizar y encontrar patrones de similitud entre usuarios de un sistema de recomendación.

información, incluso el usuario más activo puede que haya comprado o calificado menos del 1% de los ítems que se ofrecen. De acuerdo a este planteamiento, un sistema de recomendación, podría estar imposibilitado de hacer una recomendación de un ítem a un usuario en particular al no poderlo comparar con ningún otro usuario con gusto similares; teniendo como resultado, una precisión escasa.

- **Escalabilidad:** Los algoritmos de predicción y recomendación utilizados para el filtrado colaborativo implican una fuerte carga computacional la cual crece en dos direcciones: respecto al número de usuarios registrados y respecto al número de ítems disponibles para calificar.

  Al plantearse un escenario típico de un sistema de recomendación, se contará con una cantidad considerable de usuarios que solicitan recomendaciones concurrentemente; estas solicitudes exigirán del sistema un soporte suficiente para que se realicen los procesos transaccionales y los procesos de análisis y predicción.

  En un escenario de Internet, miles o millones de usuarios hacen uso de un sistema de predicción y recomendación; este crecimiento acelerado podría traer consigo problemas de escalabilidad.

CAPÍTULO 2

Sistemas de recomendaciones

2.1. ¿Qué son los sistemas de recomendación?

Los sistemas de recomendación son sistemas que tratan de ser una alternativa al proceso social de recomendación, es decir, ese acto habitual que todos practicamos al recurrir a las opiniones de conocidos o expertos cuando tenemos que tomar una decisión para adquirir algo sin tener la suficiente información para ello.

Estos sistemas actúan como guías que nos orientan para tomar decisiones relacionadas con nuestros gustos personales.

Hasta ahora las aplicaciones principales que se han hecho de estos sistemas en Internet han sido en sitios dedicados al comercio electrónico, proporcionando al usuario una forma de encontrar productos que le podría interesar adquirir.

A continuación se detalla generalidades de dos tipos de sistemas de recomendaciones de mayor uso en la actualidad: los sistemas basados en contenido y sistemas basados en filtrado colaborativo.

2.2. Sistemas de recomendación basados en el contenido

Los sistemas de recomendación basados en contenido, emplean técnicas de recuperación de información. Por ejemplo, un documento de texto es recomendado basado en una comparación entre su contenido y el del perfil del usuario.

Típicamente, el perfil muestra una lista de palabras clave y sus pesos correspondientes. Dicho perfil puede ser definido explícitamente, el usuario contesta cuestionarios, o de forma semiautomática en base a diversas heurísticas.[5]

Para identificar el tema del documento se hace un análisis de frecuencia para extraer las palabras clave. Si a un usuario le gusta un documento, los pesos de las palabras extraídas se añaden a los pesos de las palabras correspondientes en el perfil del usuario. Este proceso es conocido como retroalimentación de relevancia.[6]

Este método de recomendación presenta algunos problemas como la sobre-especialización; el sistema sólo muestra al usuario elementos similares a los que ya ha visto anteriormente. Algunas veces este problema es resuelto agregando a la búsqueda aleatoriedad (mediante algoritmos genéticos). Otro problema se presenta al encontrar información multimedia (con frecuencia presente en

[5] Lieberman, H., Selker T.: (1997) Out of Context: computer systems That Adapt to, and Learn from Context. IBM Systems Journal, Vol 39, Nos 3&4, pp. 617-631, 2000.
[6] M. Balabanovic and Y. Shoham, (1997) "Content-Based Collaborative Recommendation," Comm. ACM, Mar.1997, pp. 66-72.

páginas de Web) puesto que cuando las recomendaciones son hechas sobre documentos de texto, está información es ignorada.

Los sistemas basados en el contenido parten de una idea central: para recomendar algo a alguien es necesario que lo que le vayamos a recomendar sea muy similar (tenga un contenido muy parecido) a ciertos objetos que ya sabemos, con seguridad, son del agrado del consumidor en cuestión.

Es decir, un sistema basado en contenido, solamente recomendaría un objeto a alguien si ya conoce que su contenido tuvo éxito con él en un pasado. Esta forma de recomendar presenta un problema: el reducir las recomendaciones a unos contenidos muy similares puede hacer que el sistema tienda a entrar en una sobre-especialización con los usuarios, en otras palabras, con el tiempo quedarían muy marcados los temas de selección, siendo difícil que el sistema pudiera salir de esta tendencia si no consideramos mecanismos que puedan resolverla.

Un mecanismo utilizado en los sistemas por contenido para solucionar el problema de sobre-especialización es el introducir a las búsquedas cierta aleatoriedad. Esto consigue que de vez en cuando el sistema rompa la tendencia de recomendar siempre “los mismos temas” y pueda ampliar de esta manera el abanico en sus recomendaciones.

2.3. Sistemas de recomendación basados en un filtro colaborativo

Un sistema de recomendación basado en un filtro colaborativo podría definirse como: “aquel sistema en el que las recomendaciones se realizan basándose solamente en los términos de similitud entre los usuarios”.[7]

Estos sistemas de recomendación presentan elementos que les han gustado a otros usuarios con gustos similares, con este propósito, calculan la similitud entre usuarios. En estos sistemas, el usuario debe realizar una evaluación previa sobre algunos elementos y de esta forma se va formando el perfil del usuario.

Para cada usuario se crea un conjunto de "vecinos cercanos", usuarios cuyas evaluaciones anteriores tienen grandes semejanzas a las del usuario en cuestión. Los resultados para los elementos no calificados se predicen en base a la combinación de puntos (scores) conocidos de los vecinos cercanos.

En el filtrado colaborativo, el sistema no analiza los elementos evaluados, sino que las recomendaciones se basan solamente en la similitud entre usuarios. Esto trae consigo algunos problemas, como se comenta a continuación.

Cuando un usuario llega al sistema, no es posible hacerle recomendaciones hasta que su perfil sea lo suficientemente completo

[7] M. Balabanovic and Y. Shoham, (1997) “Content-Based Collaborative Recommendation,” Comm. ACM, Mar.1997, pp. 66-72.

para encontrarle a su grupo de vecinos cercanos. Además si los gustos del usuario son poco comunes, encontrarle un conjunto de vecinos cercanos será una tarea complicada. Esto hace notar que las recomendaciones dependen directamente del número y variedad de usuarios en el sistema.

En estos sistemas la identificación de comunidades de interés emergentes en la población de usuarios es automática, lo que permite mejoras en la conciencia de grupo y la comunicación entre éstos.[8] Estos sistemas han ganado aceptación de la gente por la ayuda que brindan en el filtrado de información.

2.4. ¿Cómo funciona un sistema de recomendación?

El funcionamiento de estos sistemas básicamente consiste en pedir al usuario que califique a una serie de preguntas que se le presentan, puntuaciones que luego el sistema cruza con las de otros usuarios con gustos similares para finalmente mostrar una serie de productos recomendados.

Los sistemas de recomendación se basan en el filtrado colaborativo de información que hace que llegue al usuario lo que podría ser más de su interés teniendo en cuenta sus gustos y preferencias. El objetivo

[8] M. Balabanovic and Y. Shoham, (1997) "Content-Based Collaborative Recommendation", Comm. ACM, Mar.1997, pp. 66-72.

de estos sistemas es encontrar la información que otros usuarios de similares características han encontrado útil y recomendarla.[9] De forma general, el término "sistema de recomendación" hace referencia tanto a los sistemas que se dedican a recomendar listas de productos como a los que ayudan a los usuarios a evaluar dichos productos.[10]

Ayudar al usuario en su proceso de decisión mediante diferentes mecanismos, pasa necesariamente por una labor de análisis de las grandes bases de datos que sustentan los sitios virtuales. Los sistemas de recomendación son una aplicación particular de descubrimiento de conocimiento en bases de datos, utilizando técnicas tales como: análisis de conglomerados, redes bayesianas, filtrado y recuperación de la información, reglas de asociación, agentes, entre otros. Estos sistemas modelan el comportamiento del usuario en base al que aplican los mecanismos para facilitarle la búsqueda de los productos y/o servicios que desea adquirir a la vez que la posibilidad de evaluación de las diferentes alternativas ofertadas.

[9] Wyner (1998), G.A. Collaborative filtering: Research or IT?. Marketing Research. Chicago. Otoño 1998. Vol 10, nº3, pp. 35-37. [C2_2]
[10] Schafer, J.B., Konstan, J. & Riedl, J. (2000), Electronic Commerce Recommender Applications. Journal of Data Mining and Knowledge Discovery. Vol. 5, nº 1/2, pp. 115-152.

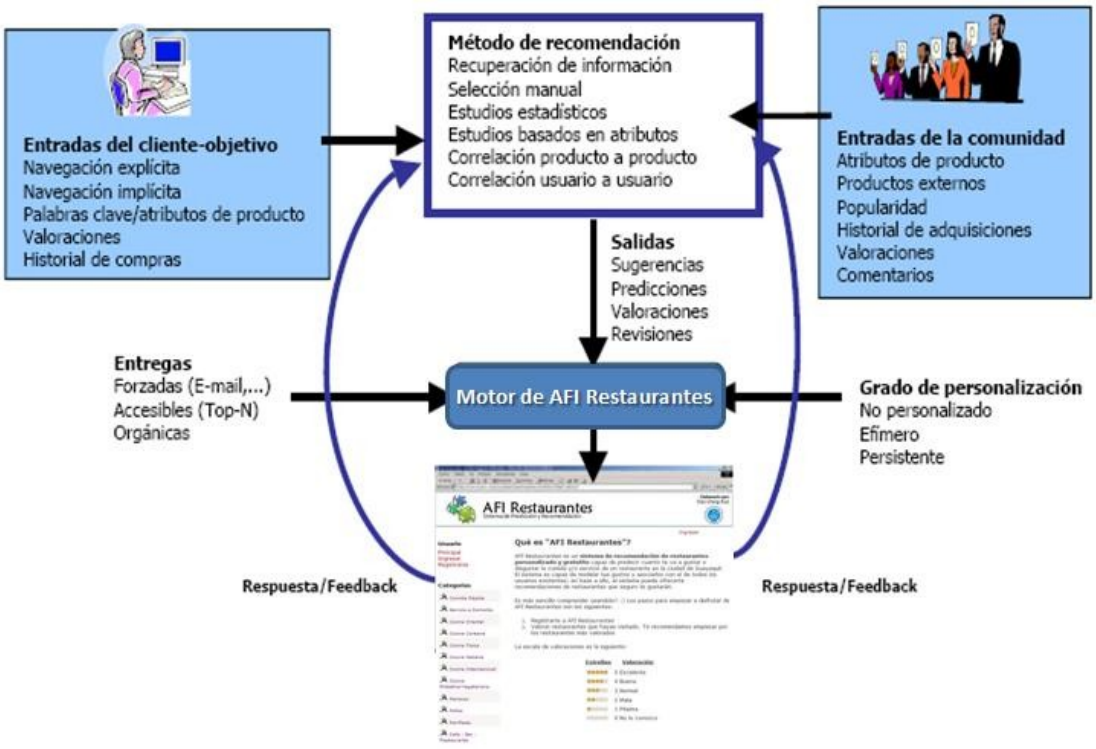


Figura 2 - Taxonomía de los sistemas de recomendación[11]

Los primeros pasos en el filtrado colaborativo vinieron de la mano de Xerox PARC, de su sistema Tapestry (ref. bibliográfica C2-4). Surgirían entonces distintos proyectos y sistemas de recomendación de productos. Ejemplos son GroupLens (ref. bibliográfica C2-5), Ringo (ref. bibliográfica C2-6), EachMovie (ref. bibliográfica C2-7) y el incluido en Amazon.com (ref. bibliográfica C2-8).

[11] Schafer, J.B., Konstan, J. & Riedl, J. (2000), Electronic Commerce Recommender Applications. Journal of Data Mining and Knowledge Discovery. Vol. 5, nº 1/2, pp. 115-152.

2.5. Mecanismos de aprendizaje de los sistemas de recomendación

Es evidente que los sistemas de recomendación deben poseer conocimiento o aprendizaje de los datos que analizan para poder sugerir o generar recomendaciones; por lo tanto, este conocimiento es generado gracias a las técnicas de generación de conocimiento no asistido usadas en la minería de datos con el uso de técnicas y algoritmos estadísticos/probabilísticas, funciones de distancias , así como muchas más técnicas de inteligencia artificial, cuyo desarrollo en conjunto con el poder de computo alcanzado hasta ahora han ayudado a la humanidad en las últimas décadas, a manejar de manera más precisa y optima amplios orígenes de información con los que tenemos que enfrentar y sacar de ellos la más relevante a nuestro interés.

Un sistema de recomendación debe evolucionar en el tiempo en cuanto a la calidad en las recomendaciones que realiza. En otras palabras, estos sistemas deben servirse de la experiencia y de la cantidad de información que utilizan para ir mejorando sus pronósticos.

Llamamos mecanismos de retroalimentación a las técnicas utilizadas por los sistemas de recomendación para ir recopilando más información a cerca de los gustos de los usuarios existentes.

Las técnicas más utilizadas para realizar la retroalimentación en los sistemas de recomendación son el mecanismo de retroalimentación implícita y el mecanismo de retroalimentación explícita.

2.5.1. Sistemas para gestionar el conocimiento

Tras el análisis de las características que deben cumplir las herramientas para la gestión del conocimiento, proponemos clasificar por un lado las herramientas que dan más énfasis en facilitar el trabajo colaborativo para la generación de conocimiento comunitario como la red de conocimiento Saber Latino, de Special Interest Group, cuyo objetivo es facilitar un lugar de encuentro de instituciones y actores de la gestión del conocimiento en el ámbito Iberoamericano[12]; y por otro lado, las herramientas que hacen mayor énfasis en la generación de estructuras de conocimiento como KnowCat que tiene como meta la creación incremental de conocimiento estructurado.[13]

La Figura 3 - Clasificación de los sistemas para la gestión de conocimiento, muestra ejemplos de sistemas para cada clasificación existente.

[12] Diseño Visual. Redes de Conocimiento como puentes culturales. [en línea], disponible en http://www.disenovisual.com/temas/temas.php

[13] RedIRIS. KnowCat – Catalizador de conocimiento [en línea], disponible en http://www.rediris.es/rediris/boletin/58-59/ponencia2.html

Figura 3 - Clasificación de los sistemas para la gestión de conocimiento[14]

Existen también herramientas, que proveen de mecanismos de trabajo colaborativo como permiten la organización interna de la memoria común de conocimiento a las cuales hemos denominado: sistemas integrales para la gestión del conocimiento.

[14] Schafer, J.B., Konstan, J. & Riedl, J. (2000), Electronic Commerce Recommender Applications. Journal of Data Mining and Knowledge Discovery. Vol. 5, nº 1/2, pp. 115-152.

2.5.2. Sistemas enfocados al trabajo colaborativo del conocimiento.

En contraste con los anteriores, hay herramientas que ponen el énfasis en el manejo colaborativo del conocimiento, dando especial importancia al usuario y sus características, y a la comunidad de usuarios como unidad de trabajo. Éstas son las herramientas que proporcionan espacios compartidos, los sistemas de recomendación y, por último, los que están destinados al aprendizaje colaborativo.

2.5.3. Espacios compartidos

En primer lugar tenemos una serie de herramientas o sistemas que nos proporcionan una interfaz de espacio compartido donde un grupo de usuarios pueden interactuar para compartir conocimiento y crear nuevo conocimiento de manera colaborativa.

Estos sistemas típicamente ofrecen una serie de funcionalidades:

- Herramientas de comunicación: mensajería, foros de debate, charla o chat.

- Herramientas para compartir contenidos: para compartir ficheros, contactos y/o enlaces.
- Herramientas de actividades conjuntas: navegación por la Web en conjunto, dibujo, edición multiusuario y calendario en grupo.

2.6. Retroalimentación implícita en sistemas de recomendación

Podemos decir que un mecanismo de retroalimentación implícita proporciona información al sistema de recomendación a cerca de los gustos de los usuarios por determinados objetos del sistema sin que estos usuarios se den cuenta de que el sistema está alimentándose y obteniendo información de ellos relativa de los gustos[15].

Para explicar más fácilmente como podría funcionar este mecanismo de retroalimentación, consideremos el ejemplo de un sistema consistente en recomendar documentos de texto a ciertos usuarios.
Cuando un usuario concreto estuviera dedicando un tiempo elevado en la lectura de un determinado documento, el sistema podría considerar que ese documento es del interés del consumidor y asignarle una puntuación elevada. De esta manera, el usuario está siendo evaluado por el sistema sin que se esté dando cuenta.

[15] Tercera Edición del Premio NAI. SRI, Sistema Inteligente de Recomendaciones. 2003

Un problema que nos surge con este tipo de mecanismos es que al no estar diciendo al usuario directamente que ese objeto es de su interés, debemos suponer que lo es a través de otras medidas como pueden ser el tiempo que está dedicando a su lectura y las veces que solicita el acceso a este documento.

Muchas veces, el que alguien dedique mucho tiempo a algo no tiene por qué significar que sea de su interés (podría ser que a la persona le haya costado entender su contenido), es por ello, que estas medidas que el sistema utiliza puede que no sean siempre apropiadas y dependan del contexto en el que se integre nuestro sistema de recomendación.

2.7. Retroalimentación explicita en sistemas de recomendación

La retroalimentación explícita es otro de los mecanismos mediante los cuales un sistema de recomendación aprende, es decir, va conociendo los gustos de los usuarios por determinados objetos[16].

Este mecanismo se basa en la acción directa por parte del usuario para indicar qué objetos determinados del sistema son de su agrado.

[16] Tercera Edición del Premio NAI. SRI, Sistema Inteligente de Recomendaciones. 2003

En el ejemplo de las recomendaciones de documentos de texto, esta técnica podría consistir en hacer que el usuario vote un determinado documento que leyó.

Otra forma más directa de que el usuario brinde información acerca del documento que leyó podría ser que simplemente, diga si le gustó ó no, pasando a engrosar la lista de documentos de su interés en ese caso ó a la lista de documentos que no lo son en el caso contrario.

CAPÍTULO 3

Análisis del proyecto

3.1 Requerimientos del sistema

- **Información de calificaciones:** son la fuente principal de información para el sistema. Los usuarios calificarán los restaurantes listados y disponibles por medio del sitio Web "AFI Restaurantes".
- **Cuentas de usuario**: Incluirán información básicas del individuo (ver 0.0.1 Modelo lógico de la base de datos). Las recomendaciones son personalizadas y deben ir relacionadas directamente con la cuenta de usuario del individuo
- **Información de establecimientos:** Información de detalle e interés sobre los restaurantes de Guayaquil: Nombre, dirección, teléfono y categoría.
- **Categorización de los establecimientos:** Los restaurantes deben ser clasificados de acuerdo al tipo de comida que brindan, con el

fin de mejorar la manera en que se realizan valorizaciones en el sitio Web.

- **Soporte para cadenas de restaurantes:** El sistema debe soportar la existencia de cadenas de restaurantes donde existe más de un local con la misma razón social.

3.2 Alcances del proyecto

- "AFI Restaurantes" se enfocará en el beneficio exclusivo del usuario, de modo que un usuario podrá saber de antemano qué establecimiento le puede gustar visitar, justificando su decisión en las valoraciones que otras personas han colaborado en realizar.
- La tarea de valoración de los establecimientos se realizará tomando una medida discreta de valores para calificar la calidad de la servicio/comida usando una escala de 0 a 5 (0 equivalente a "No lo conozco", 1 a "Pésimo" y 5 a "Excelente")
- Consideraremos diferentes criterios para la solicitud de recomendaciones. Tendremos recomendaciones generales de "Restaurantes más votados" (Top10) y tendremos las recomendaciones personalizadas, que son únicas para cada usuario, en donde podrá consultar los restaurantes en las diversas categorías que han sido determinadas.

- El sistema mostrará por cada restaurante información de interés para el usuario como la dirección y los teléfonos del establecimiento.
- El usuario tendrá la capacidad de poder sugerir un nuevo restaurante que no haya sido considerado y que desee proponer a criterio de la comunidad de calificadores en el sitio Web.
- Establecimientos o locales que pertenecen a una misma cadena de restaurantes serán valorados bajo una misma razón social, independiente de su ubicación geográfica.

3.3 Especificaciones UML de análisis

3.3.1. Diagrama de casos de uso

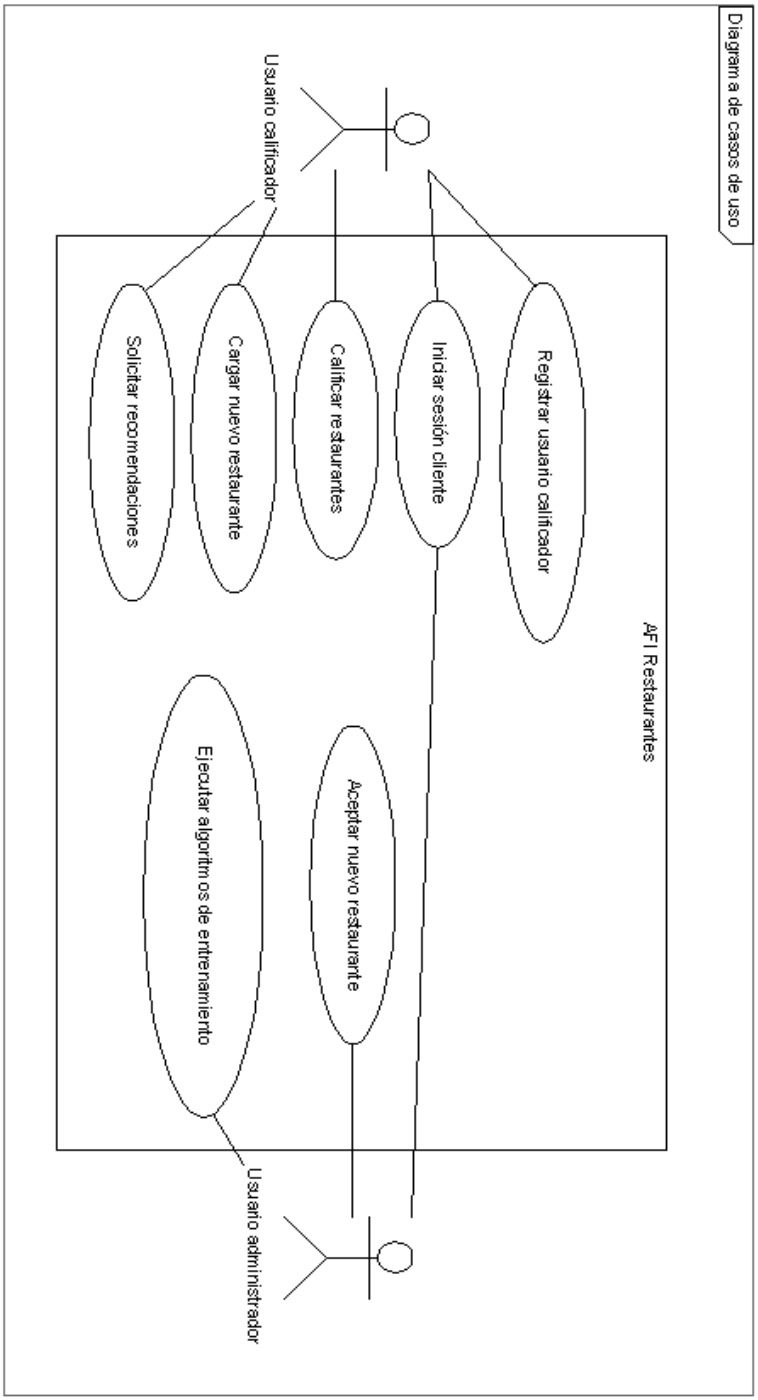


Figura 4 - Diagrama de casos de uso

3.3.1.1. Casos de uso de alto nivel

Usaremos un formato breve para la descripción inicial de los casos de uso del sistema. Esta forma de representación es muy útil para mantener una perspectiva general durante el plan de desarrollo del software.

Los casos de uso del sistema de predicción y recomendación "AFI Restaurantes" son los siguientes:

Caso de uso CU01: Registrar usuario calificador

El usuario podrá hacer uso del sistema y hacer uso de las opciones del mismo registrándose.

Caso de uso CU02: Iniciar sesión cliente

Cuando el usuario (tanto calificador como administrador) quiera navegar por el sitio, consultando recomendaciones y calificando, el sistema debe estar seguro que es un Usuario legítimo para que pueda acceder a los recursos y servicios. El cliente provee sus datos y el sistema inicia una sesión exclusiva.

Caso de uso CU03: Calificar restaurante

El usuario calificador tiene la tarea de principal de calificar los restaurantes usando una escala de 0 a 5. Esta tarea es la de

mayor interés para el sistema permitiéndole entrenarse, predecir y dar recomendaciones.

Caso de uso CU04: Cargar nuevo restaurante

El usuario calificador podría sugerir la agregación de un nuevo restaurante a la base existente.

Caso de uso CU05: Solicitar recomendaciones

El sistema presenta recomendaciones personalizadas al usuario calificador de acuerdo al criterio que éste haya especificado en el sitio.

Caso de uso CU06: Aceptar nuevo restaurante

El usuario administrador recibe las peticiones de Cargar nuevo restaurante de parte del usuario calificador y aprueba o desaprueba el nuevo ingreso.

Caso de uso CU07: Ejecutar algoritmo de entrenamiento

El usuario administrador ejecuta periódicamente los algoritmos de generación de grupos.

3.3.1.2. Casos de uso

Usaremos a continuación un formato complejo o detallado para describir cada caso de uso del sistema. Esta representación consiste en una presentación formal de casos de uso, basada en una plantilla, con campos para varias secciones de interés para la fase de análisis y diseño del proyecto. Este formato nos proveerá de mayor entendimiento sobre el significado de los casos de uso del sistema.

No existe una plantilla estándar para documentar en detalle los casos de uso. Sin embargo, existen un acuerdo en cuanto a las secciones fundamentales que deben ser consideradas en un formato detallado, las cuales hemos utilizado en la descripción de casos de uso que se muestran a continuación.[17]

[17] Wikipedia. Use Case [en línea], disponible en http://en.wikipedia.org/wiki/Use_case

Caso de Uso CU01: Registrar usuario calificador
Actor Principal: Usuario calificador. **Personal Involucrado e intereses:** ▪ **Cliente:** Quiere registrarse de una manera sencilla para poder obtener recomendaciones personalizadas. **Precondiciones:** Ingresar al sitio Web. **Poscondiciones (Garantías de éxito):** Se registró un nuevo Usuario calificador que podrá hacer uso de las opciones del sistema. **Escenario principal de éxito:** **1.** El usuario calificador solicita la creación de una nueva cuenta. **2.** El usuario calificador ingresa datos del usuario (nombre, apellido, correo electrónico, teléfono, usuario y contraseña). **3.** El sistema valida que el usuario ingresado no esté siendo utilizado por un usuario existente. **4.** El usuario calificador es registrado en la tabla de usuarios de la base de datos. **Extensiones (o flujos alternativos)** **2a.** El usuario calificador ingresa datos erróneos o incompletos. **1.** Se anula el proceso y se elimina el registro. **Requisitos especiales:** ▪ Enviar un correo electrónico confirmando el registro exitoso del usuario en el sitio Web. **Frecuencia:** ▪ Una vez por cada registro de cliente.

Tabla 1 - Caso de uso CU01: Registrar usuario calificador

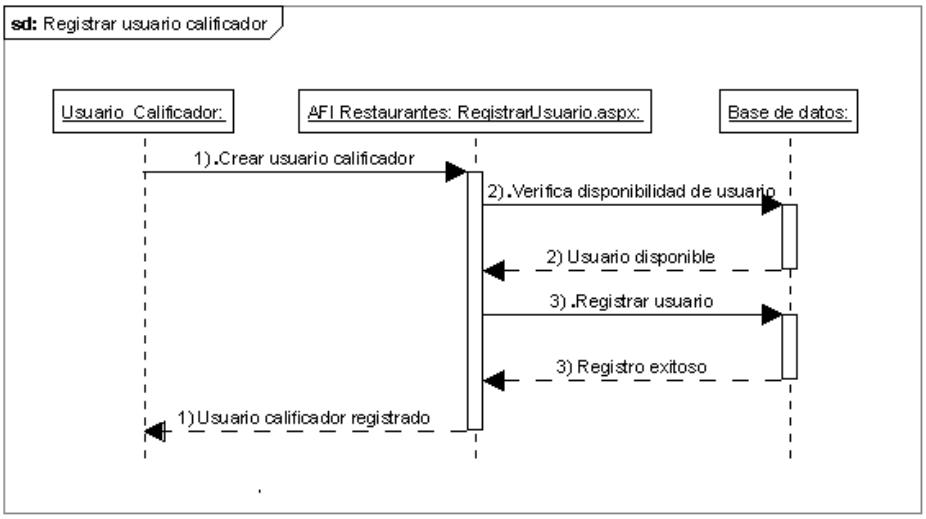


Figura 5 - Diagrama de secuencia caso de uso CU01

Caso de uso CU02: Iniciar sesión cliente
Actor Principal: Usuario calificador, usuario administrador. **Personal Involucrado e intereses:** ▪ **Usuario:** Quiere iniciar la navegación en el sitio accediendo a las distintos servicios y recursos que el sistema ofrece. **Precondiciones:** Haberse registrado en el sitio. (Caso de uso CUO1) **Poscondiciones (Garantías de éxito):** Se inició una sesión en el sistema permitiéndole navegar en él. **Escenario principal de éxito:** **1.** El usuario calificador o usuario administrador solicita el ingreso al sitio. **2.** El usuario calificador o usuario administrador ingresa los datos de usuario y contraseña. **3.** El sistema verifica que el usuario exista en la base de datos. **4.** El sistema muestra los diferentes servicios y opciones a las que puede acceder el usuario dependiendo si es administrador o calificador. **Extensiones (o flujos alternativos)** **3a.** El usuario ingresa datos erróneos o incompletos. **2.** Se anula el proceso y no acepta al usuario. **Requisitos especiales:** ▪ No existen requisitos especiales. **Frecuencia:** ▪ Cada vez que el usuario desee ingresar al sistema.

Tabla 2 - Caso de uso CU02: Iniciar sesión cliente

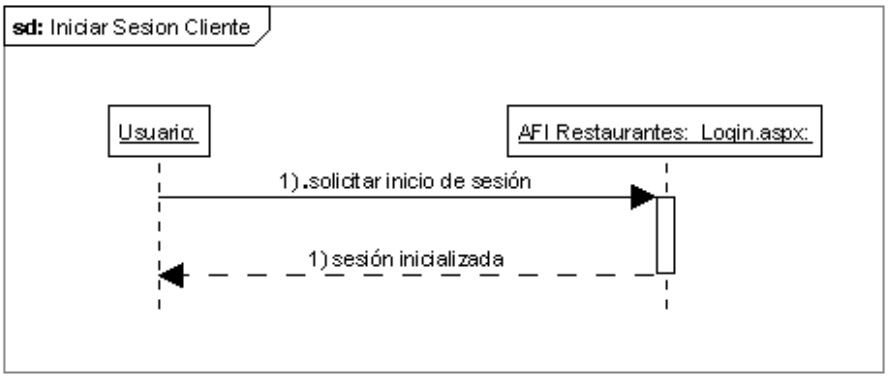


Figura 6 - Diagrama de secuencia caso de uso CU02

Caso de uso CU03: Calificar restaurante
Actor Principal: Usuario calificador. **Personal Involucrado e intereses:** ▪ **Usuario calificador:** Quiere dar una votación a un restaurante en particular. **Precondiciones:** Haber iniciado sesión. (Caso de uso CUO2) **Poscondiciones (Garantías de éxito):** Se almacena su calificación en la base de datos. **Escenario principal de éxito:** **1.** El usuario calificador identifica un restaurante conocido en la lista. **2.** El usuario calificador elige una calificación entre las opciones. **3.** El sistema almacena su valoración en la tabla de calificaciones de la base de datos. **Extensiones (o flujos alternativos)** No existen flujos alternativos en este caso. **Requisitos especiales:** ▪ No existen requisitos especiales. **Frecuencia:** ▪ Cada vez que el usuario desee calificar un restaurante.

Tabla 3 - Caso de uso CU03: Calificar restaurante

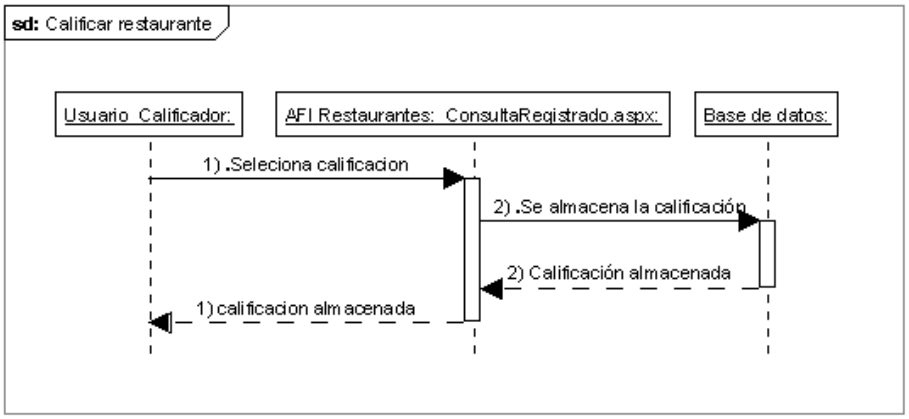


Figura 7 - Diagrama de secuencia caso de uso CU03

Caso de uso CU04: Cargar nuevo restaurante
Actor Principal: Usuario calificador. **Personal Involucrado e intereses:** ▪ **Usuario calificador:** El usuario puede sugerir la agregación de un nuevo restaurante a la base de datos. **Precondiciones:** Haber iniciado sesión. (Caso de uso CUO2) **Poscondiciones (Garantías de éxito):** Se le envía un mensaje al usuario notificándole que su solicitud será estudiada. **Escenario principal de éxito:** **1.** El Usuario calificador no identifica un restaurante conocido en la lista. **2.** El Usuario calificador solicita la incursión de un restaurante para el sitio Web. **3.** La solicitud de nuevo ingreso se pone en espera de aprobación por el usuario administrador. (Caso de uso CUO6) **Extensiones (o flujos alternativos)** No existen flujos alternativos en este caso. **Requisitos especiales:** ▪ No existen requisitos especiales. **Frecuencia:** ▪ Cada vez que el usuario desee ingresar un restaurante al sistema.

Tabla 4 - Caso de uso CU04: Cargar nuevo restaurante

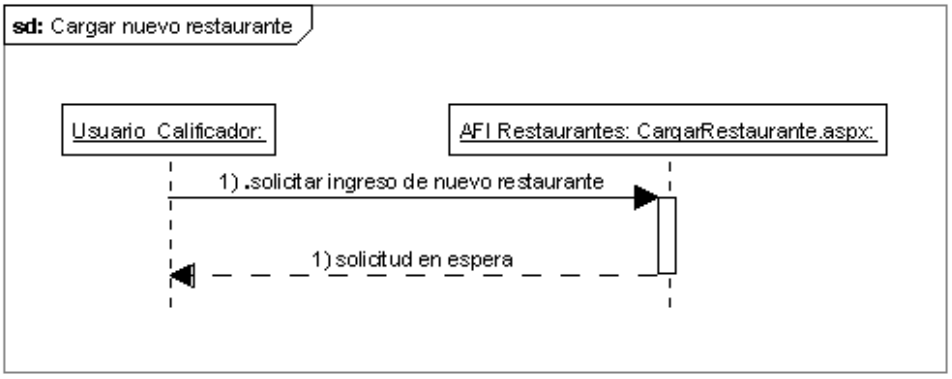


Figura 8 - Diagrama de secuencia caso de uso CU04

Caso de uso CU05: Solicitar recomendaciones
Actor Principal: Usuario calificador. **Personal Involucrado e intereses:** ▪ **Usuario calificador:** El usuario desea que el sistema le sugiera que otros restaurantes desconocidos por él le pueden gustar. **Precondiciones:** Haber calificado los restaurantes conocidos por él. (caso de uso CUO3) **Poscondiciones (Garantías de éxito):** El sistema le muestra al usuario calificador una lista de restaurantes desconocidos por él y que pueden resultar de su agrado. **Escenario principal de éxito:** **1.** El usuario calificador solicita que se le den recomendaciones. **2.** Se predice el grupo al que pertenece el usuario calificador por medio del algoritmo de Fisher. **3.** Se ejecuta el algoritmo de filtrado colaborativo sobre los usuarios del grupo asignado en el paso 2. **4.** El sistema proporciona una lista de restaurantes con una calificación (predicción) de cuanto le va a gustar dicho restaurante. **Extensiones (o flujos alternativos)** **5a.** Las recomendaciones dadas por el sistema están sujetas a la cantidad y calidad de los datos ingresados por todos los usuarios del mismo. **Requisitos especiales:** ▪ No existen requisitos especiales. **Frecuencia:** ▪ Cada vez que el usuario solicite recomendaciones al sistema.

Tabla 5 - Caso de uso CU05: Solicitar recomendaciones

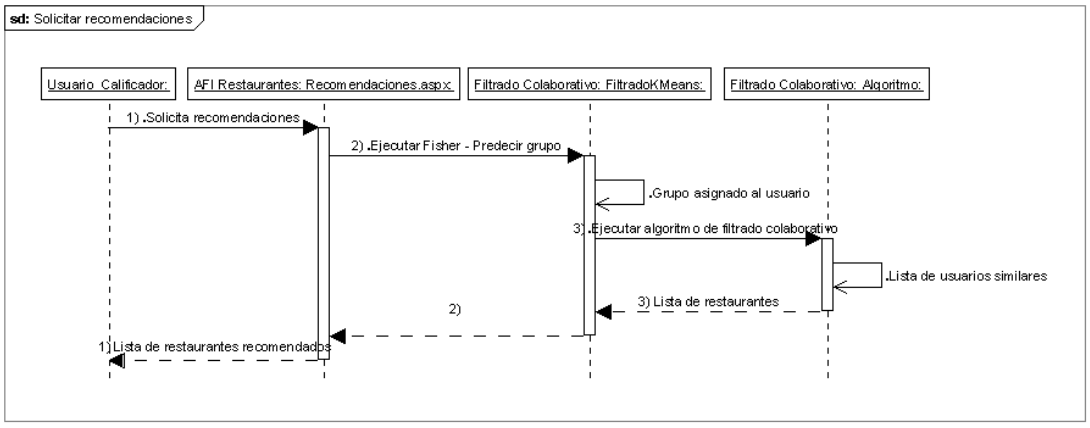


Figura 9 - Diagrama de secuencia caso de uso CU05

Caso de uso CU06: Aceptar nuevo restaurante
Actor Principal: Usuario administrador. **Personal Involucrado e intereses:** ▪ **Usuario administrador:** El usuario administrador verifica si es factible el introducir un nuevo restaurante a la base de datos. **Precondiciones:** Haber iniciado sesión. (Caso de uso CUO2) **Poscondiciones (Garantías de éxito):** Se almacena o no el nuevo restaurante en la base de datos y se le envía un mensaje al usuario calificador dándole la decisión del administrador **Escenario principal de éxito:** **1.** El usuario administrador encuentra una solicitud de agregación de un nuevo restaurante. **2.** El usuario administrador busca si ese restaurante se encuentra ya en la base de datos. **3.** El usuario administrador verifica la veracidad de la información (en la medida de lo posible). **4.** El usuario administrador agrega el nuevo restaurante a la tabla de restaurantes en la base de datos. **Extensiones (o flujos alternativos)** **2a** Si el restaurante ya existe se le envía una notificación al usuario calificador de que así es. **3a** Si la información se verifica que es falsa se detiene el proceso. **Requisitos especiales:** ▪ Enviar un correo electrónico confirmando el ingreso o no del restaurante. **Frecuencia:** ▪ Cada vez que el usuario administrador encuentre una nueva solicitud de ingreso de nuevos restaurantes.

Tabla 6 - Caso de uso CU06: Aceptar nuevo restaurante

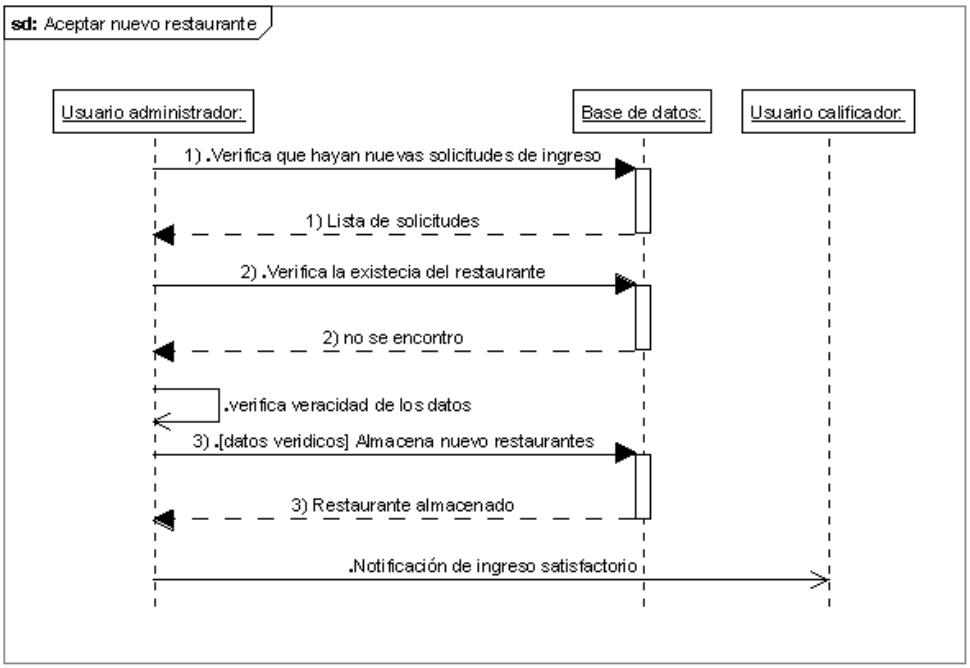


Figura 10 - Diagrama de secuencia caso de uso CU06

Caso de Uso CU07: Ejecutar algoritmo de entrenamiento
Actor Principal: Usuario administrador. **Personal Involucrado e intereses:** ▪ **Usuario administrador:** El usuario administrador ejecuta el algoritmo de k-medias para la generación de grupos de usuarios. **Precondiciones:** Haber iniciado sesión con privilegios de administrador. **Poscondiciones (Garantías de éxito):** Se almacena en la base de datos la información de los nuevos grupos formados. **Escenario principal de éxito:** **4.** El usuario administrador identifica que la variabilidad de los datos ha cambiado. **5.** El usuario administrador ingresa al módulo de Clusterización del sitio Web. **6.** Se envía el requerimiento de generación de grupos al núcleo del sistema. **7.** Se ejecuta el algoritmo de k-medias. **8.** Se almacenan los grupos generados en la tabla de grupos de la base de datos. **Extensiones (o flujos alternativos)** No existen flujos alternativos en este caso. **Requisitos especiales:** ▪ No existen requisitos especiales. **Frecuencia:** ▪ Periódicamente, de acuerdo a la variabilidad de los datos.

Tabla 7 - Caso de uso CU07: Ejecutar algoritmo de entrenamiento

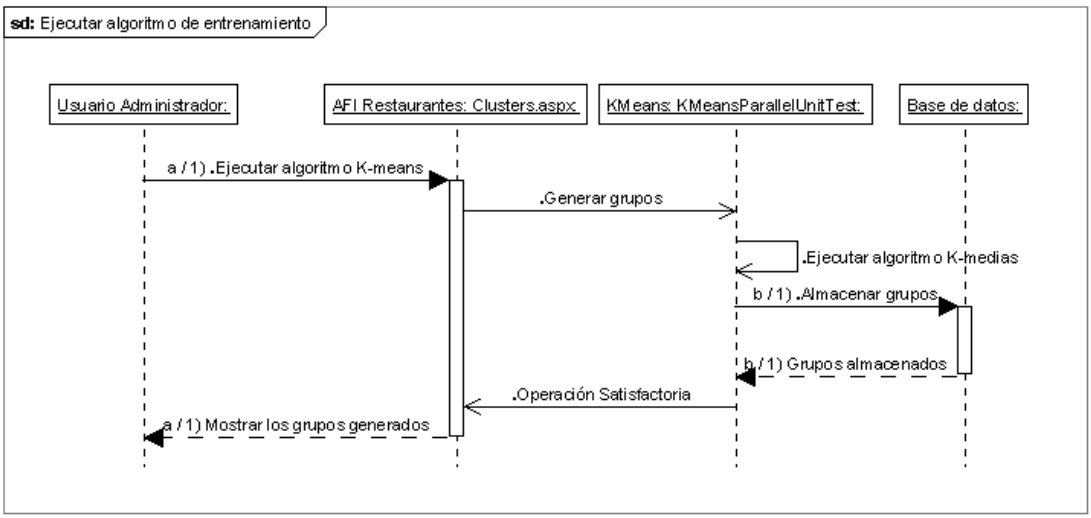


Figura 11 - Diagrama de secuencia caso de uso CU07

3.3.2. Diagrama de clases

El siguiente diagrama (ver Figura 12 - Diagrama de clases) muestra, de manera simplificada, las clases utilizadas por el sistema de recomendaciones para las tareas de predicción, generación de grupos (clusterización), elaboración personalizada de recomendaciones (filtrado colaborativo), clases para el manejo del sitio Web “AFI Restaurantes” y la administración de usuarios.

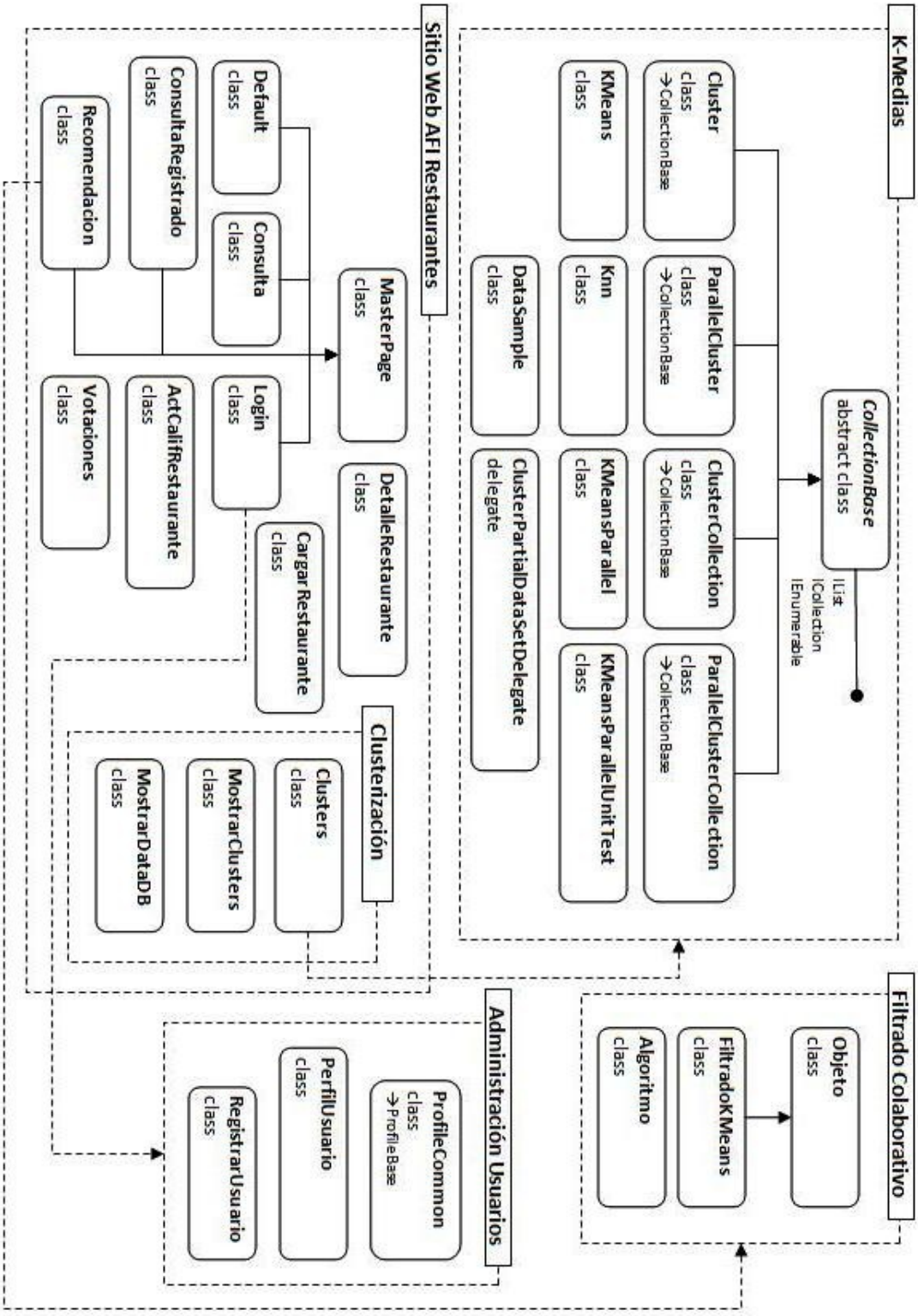


Figura 12 - Diagrama de clases

3.4 Justificaciones

3.4.1 Justificación de la técnica de minería de datos escogida

Con el fin de resolver el problema planteado, dado el modelo de extracción de conocimiento que utilizamos se usan dos tareas de minería de datos:

1. **Clasificación**: Esta tarea se utiliza con el propósito de obtener temporalmente una lista ordenada, de los usuarios más "similares" entre sí. Tarea que se va a solucionar usando el método de *vecinos más próximos o KNN (nearest neighbor)*.[18]
2. **Agrupamiento:** Esta tarea se utiliza para generar y almacenar, persistentemente, grupos de usuarios con características similares con el fin de abreviar la carga computacional, reduciendo de esta manera la cantidad de información procesada. El método que escogimos para determinar estos grupos es el de K-medias.[18]

Hasta ahora hemos hablado de tareas y método de una manera muy breve para tener un marco de referencia para el entendimiento de lo expuesto en esta sección. En el capítulo 5, en la sección 5.3.2 Tareas y métodos, se hablará de una forma

[18] Hernández, José. Introducción a la Minería de Datos. Pearson. 2004

más detallada acerca de las distintas tareas y métodos en la minería de datos, así como del modelo a seguir por nuestro sistema de recomendación.

Con el objetivo de analizar qué tareas y métodos escogeríamos para el desarrollo del presente proyecto, utilizamos la **expresividad** como nuestra medida de evaluación. La expresividad es la característica que diferencia a los métodos de aprendizaje, es la forma en cómo se expresan los patrones aprendidos. Este aspecto es fundamental, ya que determina la razón por la que ciertos métodos resultan mejor para resolver unos problemas que otros.

En la práctica, cada método permite expresar de mejor manera ciertos tipos de patrones, por eso se explica que existan tantos métodos y esto nos permite capturar ciertos tipos de patrones, si uno falla se puede probar con otro. La Figura 13 - Representación de patrones, representa tipos de patrones que se pueden obtener en base a distintos métodos.

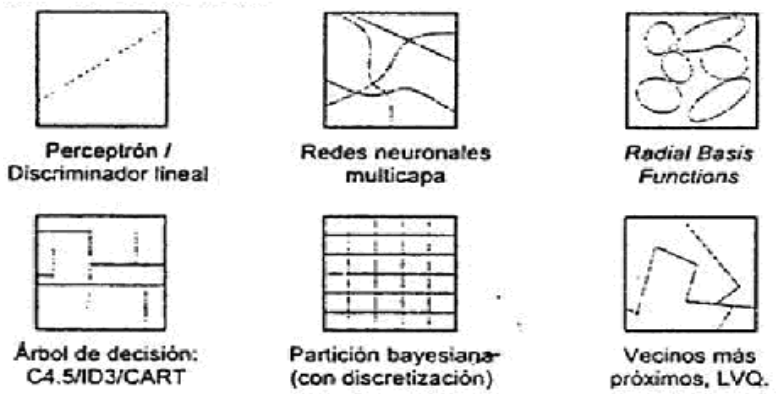


Figura 13 - Representación de patrones

Estos métodos se basan en colocar fronteras, linderos o confines entre zonas, cada una de las zonas es capaz de aglutinar un grupo. Se conocen como "cerca y rellena" (*fence* y *fill*). Otros métodos no establecen fronteras explícitamente, sino que se basan en el concepto de "centros" y clasifican/agrupan por la distancia a estos centros o zonas más densas; como por ejemplo, el método de funciones de base radial o los vecinos más próximos. Pero, de todas maneras se los conoce también como métodos de "cerca y rellena", ya que los individuos que están próximos suelen ser de la misma clase.[19]

[19] Hernández, José. Introducción a la Minería de Datos. Pearson. 2004

3.4.2 Justificación de los algoritmos escogidos

En la implementación de "AFI Restaurantes", nuestro sistema de predicción y recomendación se han introducido los siguientes algoritmos:

1. **Algoritmo de Vecinos más cercanos:**

   Este algoritmo nos permite conseguir una lista ordenada de los usuarios más similares al usuario al que se desea entregar una recomendación. Implica una tarea de comparación de un usuario contra todos, el cual representa un procedimiento de alta demanda computacional y usaremos las técnicas que se mencionan a continuación para complementar el trabajo de recomendación. Un ejemplo de aplicación se verá más adelante en la sección 3.4.3 Ejemplo de aplicación de los algoritmos escogidos.

2. **Algoritmo de K-medias:**

   Este algoritmo genera grupos de usuarios con características similares, dichos grupos se mantendrán almacenados en la base de datos para posterior uso y además formarán parte de un histórico de grupos que será almacenado en nuestro

modelo multidimensional. (Ver sección 4.4.2 Modelo multidimensional de la base de datos)

La finalidad de usar el algoritmo de K-medias es reducir la carga computacional al momento de dar recomendaciones y de resolver el problema de escalabilidad mencionado en el capítulo introductorio.

Para dicho efecto, el algoritmo K-medias hará uso de los siguientes procedimientos que mejoran su desempeño:

El test F de reducción de variabilidad

Al hacer uso del algoritmo de K-medias es necesario fijar el número de grupos que se van a generar. Este número no se puede estimar con un criterio de homogeneidad porque de ser así se formaría un solo grupo; por lo tanto, usamos el test F para estimar el número de grupos óptimo que se necesita crear. El test F le da un carácter dinámico a la generación de grupos de acuerdo a la variabilidad de los datos.

Algoritmo de Fisher

Realiza la tarea de predicción. Este procedimiento permite asignar un usuario a un determinado grupo que ha sido creado con anterioridad. Al predecir a qué grupo un usuario será asignado logramos que la tarea de

recomendaciones mejore considerablemente en su tiempo de respuesta. El algoritmo KNN (recomendación) se ejecutará comparando el usuario contra su grupo asignado y sus centroides; y ya no contra todos los usuarios registrados en el sistema.

3.4.3 Ejemplo de aplicación de los algoritmos escogidos

En un principio, consideraremos un número pequeño de usuarios que usan "AFI Restaurantes". En este escenario, el Usuario A pide una recomendación y el sistema hace uso del algoritmo de vecino más próximo para generarla. Debemos tener en cuenta que, para al momento, no existen grupos de usuarios definidos, de manera que el proceso involucra todos los usuarios existentes.

Conforme se incrementa el número de usuarios, el tiempo necesario para generar las recomendaciones para el Usuario A aumentará. Para solucionar este problema de escalabilidad, el Administrador de "AFI Restaurantes" ejecuta periódicamente el algoritmo k-medias, que en su procedimiento usa el test F, para construir la cantidad óptima de grupos entre los usuarios existentes. Los grupos son almacenados en la base de datos.

La próxima vez que el Usuario A solicite una recomendación, se usa el algoritmo de Fisher para predecir a qué grupo pertenece, y se prosigue a usar el algoritmo de vecino más cercano sobre los usuarios de dicho grupo, ya no involucrando a todos los usuarios existentes.

CAPÍTULO 4

Diseño del proyecto

4.1 Diseño arquitectónico

4.1.1. Definición de la estructura del sistema

"AFI Restaurantes", nuestro sistema de predicción y recomendaciones de restaurantes de Guayaquil está dividida en tres capas:

- **Presentación**: Son las páginas ASPX del sitio Web que serán vistas en el explorador.
- **Lógica de aplicación**: Son las tareas y los métodos de minería de datos que se utilizan al operar con el sistema "AFI Restaurantes".
- **Datos**: usaremos el sistema de base de datos Microsoft SQL Server 2005 Express Edition.

Utilizaremos las herramientas de ASP.NET 2.0 de Microsoft Visual Studio .NET 2005, como plataforma de desarrollo para nuestra

aplicación Web. Usaremos Poseidon UML y Visio 2003 para la elaboración de los diagramas.

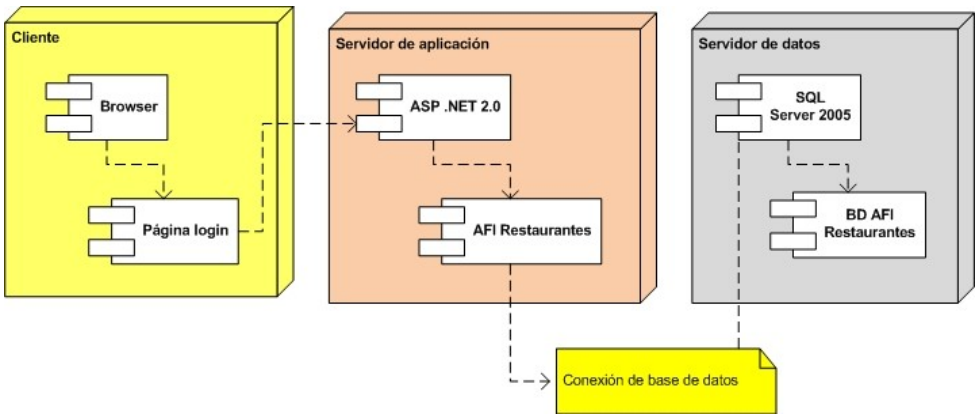


Figura 14 - Diseño arquitectónico de "AFI Restaurantes"

La Figura 14 - Diseño arquitectónico de "AFI Restaurantes" muestra la estructura del sistema; a continuación se expone los objetivos en cada nivel de esta división.

4.1.2. Objetivo de cada capa

- **Nivel de presentación:** Representa la interacción con el cliente que usa un navegador Web. Sirve como interfaz para las entradas del cliente (datos personales de registro, ingreso de calificaciones) y para las salidas del sistema (lista de recomendaciones, etc.).

El objetivo de este nivel es lograr que el cliente tenga facilidad en la navegación, en el ingreso de calificaciones y solicitud de recomendaciones en un ambiente visualmente atractivo.

- **Nivel de aplicación:** Representa los procedimientos y operaciones que se realizan para responder a las peticiones del cliente.
- **Nivel de datos:** Representa los datos adquiridos de la interacción entre el sistema y los usuarios. El objetivo de esta capa es servir de repositorio y brindar acceso a la información necesaria en el momento requerido.

4.1.3. Ventajas y desventajas de la arquitectura

Ventajas

- Aísla la lógica de la aplicación en componentes independientes que podrían reutilizarse.
- Se realiza poco procesamiento en la capa de presentación y las peticiones de trabajo se envían al nivel intermedio.
- Distribución de las capas en varios nodos físicos de cómputo y en varios procesos. Esto mejora el

desempeño, la coordinación y el compartir información en un sistema de tipo cliente-servidor.

Desventajas

- No se puede representar la lógica en componentes aislados.
- Se deben tener precauciones en la capa de nivel intermedio debido a la carga de procesamiento al que se encuentra expuesta.

4.2 Especificaciones UML de diseño

4.2.1 Interacción entre capas

Los paquetes de presentación se comunican con la capa de aplicación a través de ASP.NET 2.0, enviando sus peticiones usando el protocolo HTTP. De la misma manera, cuando la aplicación responde usa el mismo protocolo.

La capa de aplicación interactúa con la capa base estableciendo una conexión directa con la base de datos de "AFI Restaurantes".

4.2.2 Modelo de componentes

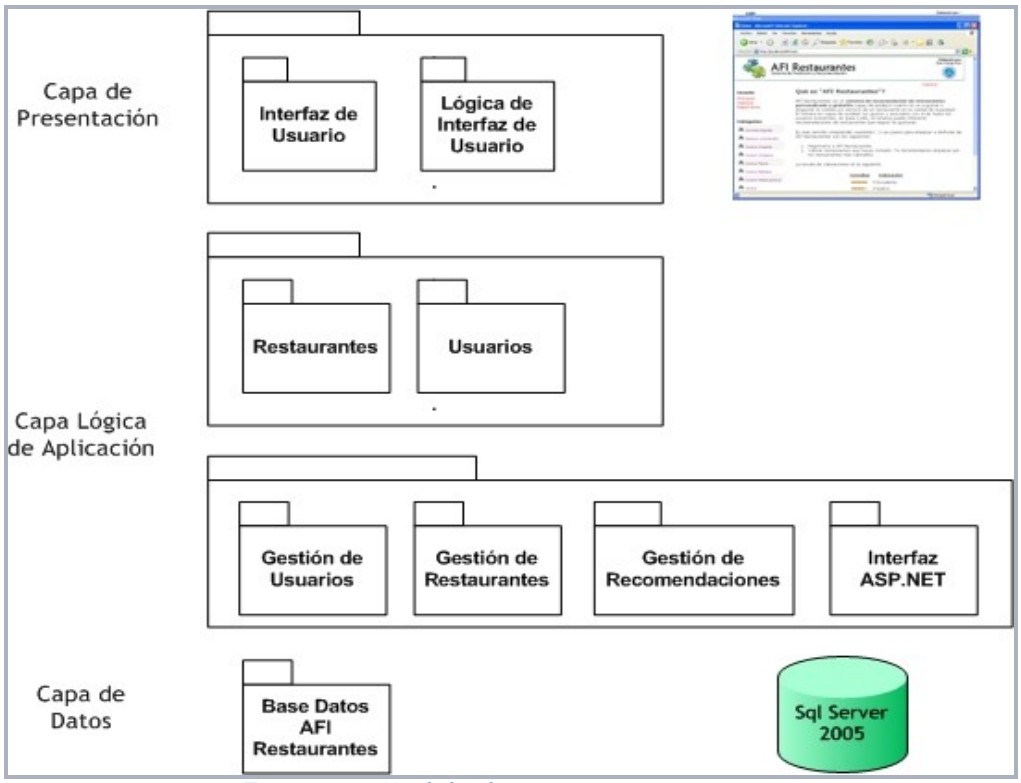


Figura 15 - Modelo de componentes

4.3 Flujo de ventanas y layouts

En el sistema hemos definido dos tipos de ambientes de trabajo: las tareas que son realizadas por el usuario calificador y las tareas realizadas por el usuario administrador. De acuerdo a dicho planteamiento tendremos el siguiente flujo de ventanas:

4.3.1 Flujo de ventanas para el usuario administrador

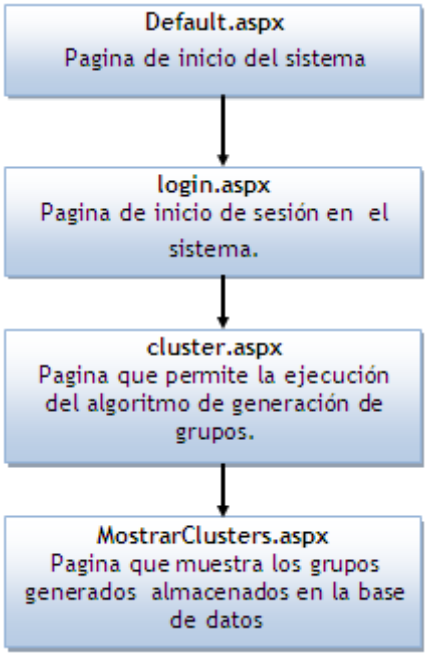


Figura 16 - Flujo de ventanas (Usuario administrador)

4.3.2 Flujo de ventanas para el usuario calificador

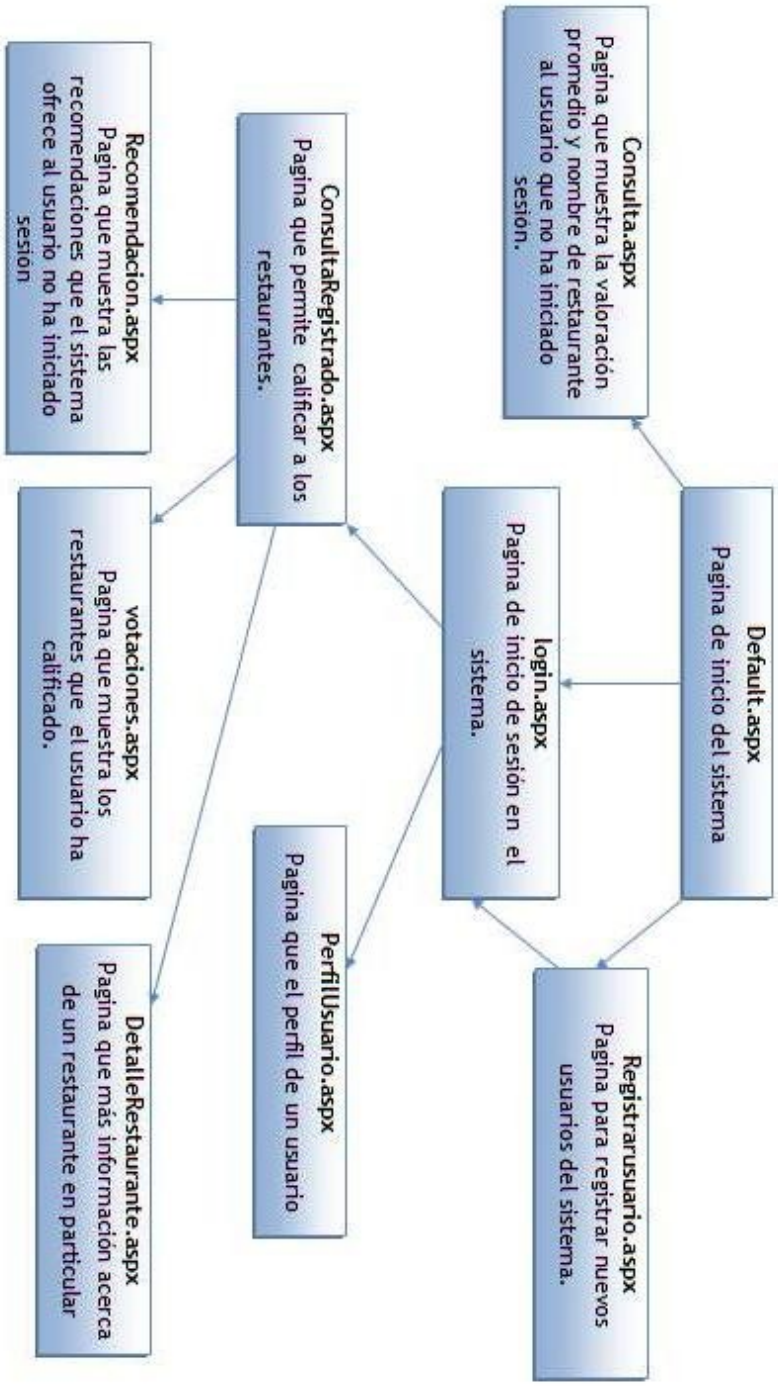


Figura 17 - Flujo de ventanas (Usuario calificador)

4.4 Modelo de la base de datos

4.4.1 Modelo lógico de la base de datos

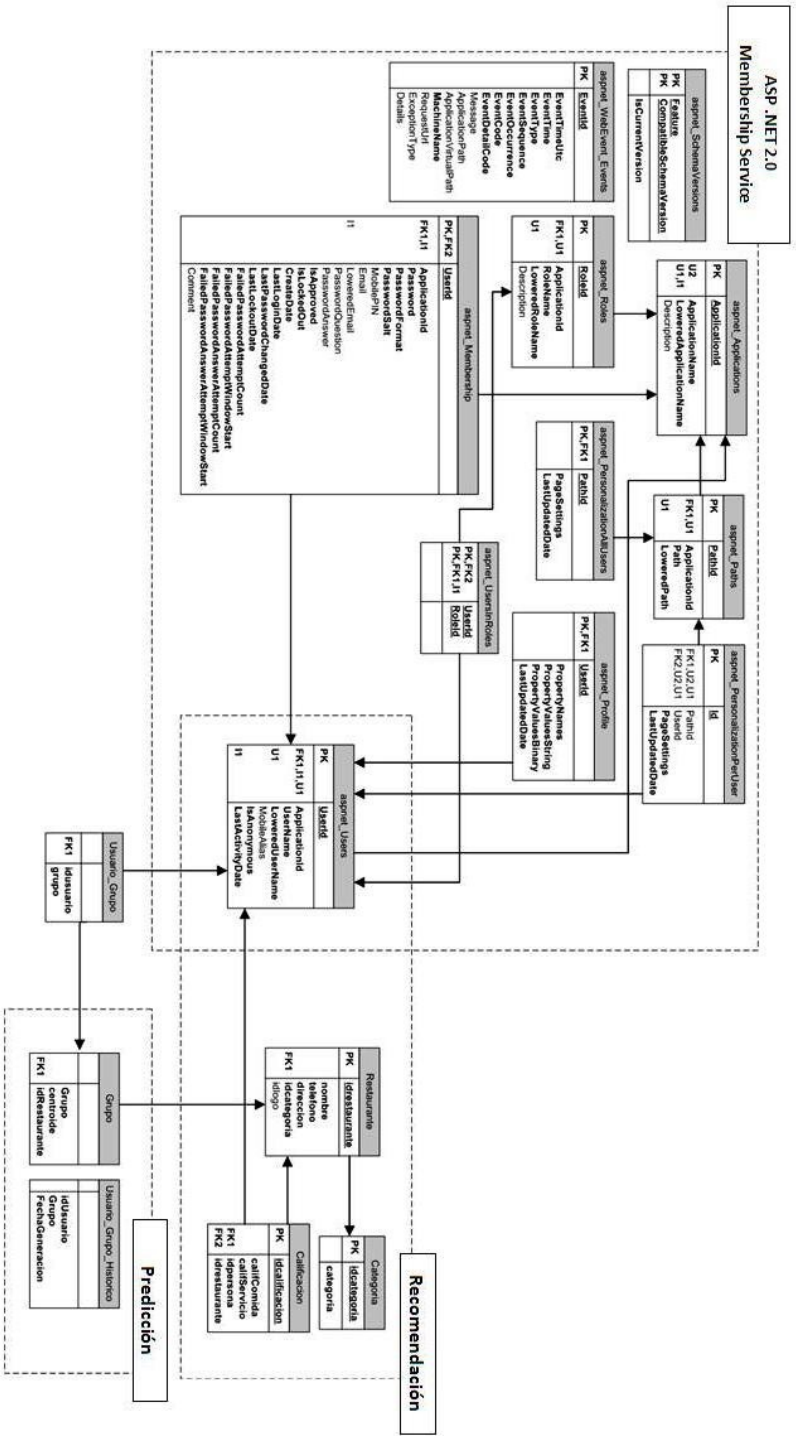


Figura 18 - Modelo lógico de base de datos

4.4.2 Modelo multidimensional de la base de datos

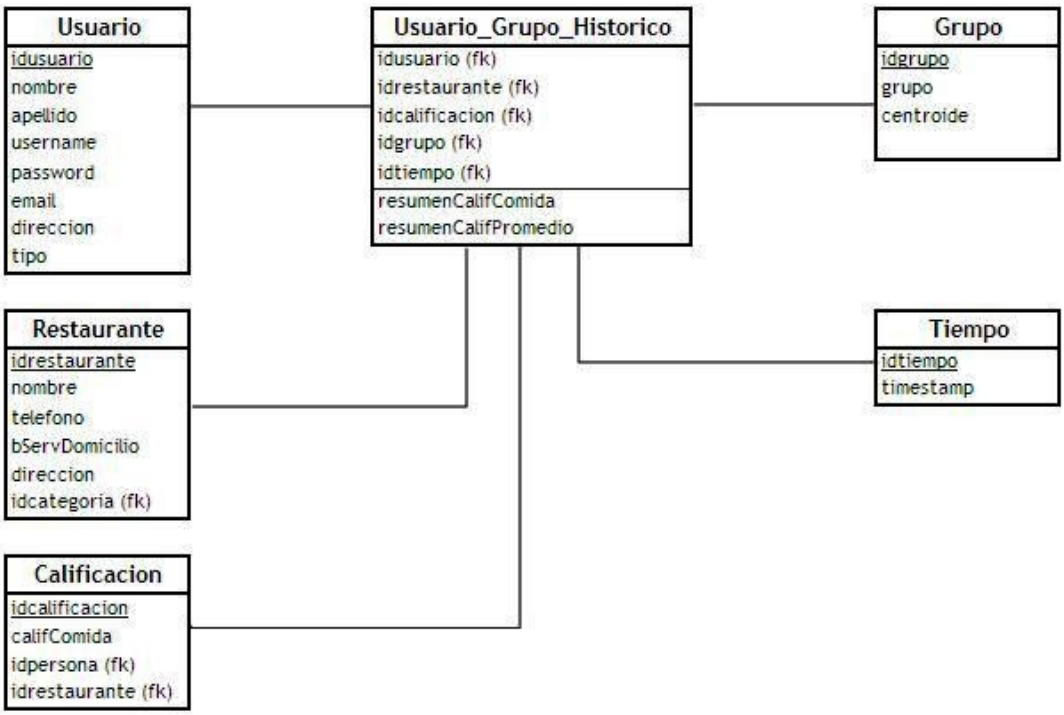


Figura 19 - Modelo multidimensional de base de datos

CAPÍTULO 5

Implementación del proyecto

5.1 Círculo virtuoso de la minería de datos

Se considera como un círculo virtuoso al conjunto consecutivo de fases que intervienen en el proceso de minería de datos a una base de datos, los pasos a seguir para la realización de un proyecto de minería de datos son casi siempre los mismos, independientemente de la técnica específica de extracción de conocimiento usada.

El proceso de minería de datos pasa por las siguientes fases:

a) Filtrado de datos

b) Selección de variables

c) Extracción de conocimiento

d) Interpretación y evaluación

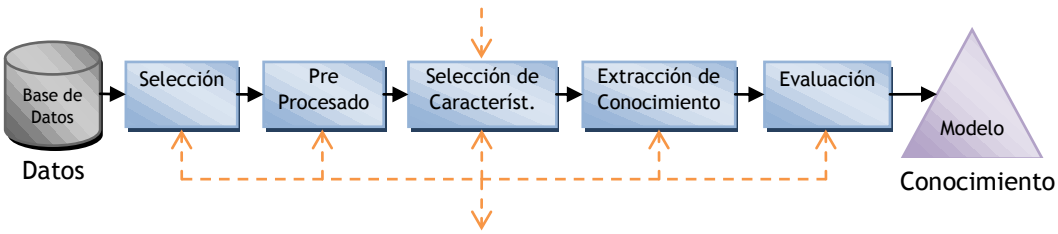


Figura 20 - Fases de la minería de datos[20]

[20] DAEDALUS Data. Minería de Datos. [En línea], disponible en: www.daedalus.es/AreasMDFase4-E.php

a) Filtrado de datos

En el sistema "AFI Restaurantes" no es necesario realizar la tarea de filtrado de datos para realizar las tareas de minería de datos debido a que el sistema se alimenta de forma explícita, en otras palabras la información necesaria es introducida directamente por el usuario al momento de dar sus calificaciones. Estas calificaciones están sujetas a un rango específico por lo cual también se descarta el hecho de tener datos aberrantes. La única situación adversa se presentaría si el usuario decide voluntariamente escoger calificaciones que no sean de su real opinión, bajo estas circunstancias las recomendaciones no serán de fiabilidad.

b) Selección de variables

En nuestro proyecto de tesis al contar con una alimentación explicita y un patrón definido, la selección de las variables que participan en el proceso está definida con anterioridad por lo tanto esto no representa un problema.

Las variables que son de más importancia en nuestra solución son el usuario, el restaurante, la calificación que el usuario le da al restaurante y el grupo al que el usuario pertenece, siendo esta ultima una variable que el sistema genera.

c) Extracción de Conocimiento

Se usan dos técnicas en el sistema "AFI Restaurantes" para representar los patrones de comportamiento observados en los valores de las variables del problema. La primera es la de agrupamiento para generar grupos de usuarios "afines" entre sí. La segunda es la de clasificación para obtener de un grupo específico una lista ordenada de los usuarios más "similares" con el propósito de obtener de ellos una lista de restaurantes para las recomendaciones.

d) Interpretación y evaluación

Finalmente se procede a su validación, comprobando que las conclusiones son válidas y satisfactorias. La interpretación del resultado así como su evaluación está determinada por la aceptación del usuario hacia la recomendación y predicción ofrecida por el sistema.

5.2 Análisis de clúster (o Análisis de conglomerados)

El análisis clúster, también conocido como análisis de conglomerados, taxonomía numérica o reconocimiento de patrones, es una técnica estadística multivariante cuya finalidad es dividir un conjunto de objetos en grupos de forma que los perfiles de los objetos en un mismo grupo sean muy similares entre sí (cohesión interna del grupo)

y los de los objetos de clústeres diferentes sean distintos (aislamiento externo del grupo)[21].

El Análisis Clúster tiene como propósito esencial, agrupar aquellos objetos que reúnan idénticas características, es decir, se convierte así en una técnica de análisis exploratorio diseñada para revelar las agrupaciones naturales dentro de una colección de datos.

Esta técnica es usada por el sistema "AFI Restaurantes" para generar grupos de usuarios similares, y procesar al grupo más conveniente, obteniendo de esta manera una solución al problema de escalabilidad.

5.2.1 Fundamentos del análisis de clúster

El procedimiento utilizado en esta técnica es la de elaborar grupos de usuarios mediante las calificaciones otorgadas a los restaurantes, cabe recalcar que con el fin de obtener unos centro de agrupamiento o centroides que sirvan para futuras referencias se escogen a todos los restaurantes que se encuentren almacenados en su respectiva tabla en la base de datos, se encuentren calificados o no.

[21] Salvador Figueras, M (2001): Análisis de conglomerados o Clúster" [en línea], http://www.5campus.org/leccion/cluster

Los principios fundamentales implicados en cualquier Análisis de Clúster se muestran en la Tabla 8 - Fundamentos del análisis de clúster.

Informe de Aglomeración	**Ofrece información sobre los objetos o casos que se combinan en cada etapa de un proceso de agrupación jerárquica.**
Centroides de Agrupamiento	Son los valores medios (medias) de las variables para todos los casos u objetos de un grupo particular.
Centros de Agrupamiento	Son los puntos de partida iniciales en la agrupación no jerárquica. Los grupos se construyen alrededor de estos centros o semillas.
Participación en el Grupo	Indica el grupo al que pertenece cada objeto o caso.
Dendrograma	Llamado también gráfica de árbol, es un dispositivo gráfico para presentar los resultados del conglomerado. Las líneas verticales representan los grupos que están unidos. La posición de la línea en la escala indica las distancias en las que se unieron los grupos. Se lee de izquierda a derecha.
Distancias entre Centros de Grupos	Indican la separación existente entre los pares individuales de los grupos. Los grupos muy separados son distintos y, por tanto, deseables.
Diagrama de Carámbanos	Es una representación gráfica de los resultados del conglomerado, se llama así porque se asemeja a una hilera de carámbanos que pende del alero de una casa. Las columnas corresponden a los objetos que se agrupan y los renglones corresponden al número de conglomerados. Se lee de abajo hacia arriba.
Matriz de Coeficientes de Distancia/Similitud	Es una matriz de triángulo inferior que contiene las distancias en dirección pareada entre los objetos o casos.

Tabla 8 - Fundamentos del análisis de clúster[22]

5.2.2 Métodos básicos de partición: Algoritmo K-medias

Este tipo de método es conveniente utilizarlo cuando los datos a clasificar son muchos y/o para refinar una clasificación

[22] Nores, José Emilio Gondar. Artículos estadísticos – Análisis Clúster [En línea], www.estadistico.com/arts.html?20001023-1

obtenida utilizando un método jerárquico. Supone que el número de grupos es conocido a priori.

La forma como lo implementamos en nuestra solución es la siguiente:

Se comienza con la elaboración arbitraria de un número específico de grupos (dos en nuestro caso), donde los usuarios se van a ir ubicando al grupo que le corresponde con cada iteración, en la cual se verifica su cercanía con los centroides da cada grupo, y luego el centroide es re calculado. El número de grupos también se va ir modificando (en este caso incrementalmente) de acuerdo a la prueba F de reducción de variabilidad que es usada justamente para determinar la cantidad adecuada de clústeres.

5.2.3 Método jerárquico: Algoritmo KNN o vecino más cercano

5.2.3.1 *Métodos jerárquicos*

Se caracterizan porque en cada paso del algoritmo sólo un objeto cambia de grupo y los grupos están anidados en los de pasos anteriores. Si un objeto ha sido asignado a un grupo ya no cambia más de grupo.

5.2.3.2 *Algoritmo KNN (vecino más cercano)*

El algoritmo KNN es un algoritmo que permite el descubrimiento de conocimiento no asistido y pertenece a los métodos sin modelo, y retardados o perezoso (*lazy*) que se basa en actuar para cada pregunta o predicción requerida.

En la solución planteada para nuestro proyecto de tesis nos ayuda a determinar los usuarios más cercanos al cual le queremos ofrecer una recomendación, mediante las calificaciones que los usuario han hecho a los distintos restaurantes que se encuentran almacenados en nuestro sistema.

5.3 El problema de extracción de patrones

5.3.1 Introducción

Se dice comúnmente que el proceso de minería de datos convierte datos en conocimiento. En algunos casos, hasta se llega a decir que el objetivo es extraer "verdad a partir de basura" [Thornton 2000]. La perspectiva sobre nuestro proceso de extracción de patrones se podría resumir como se ilustra en la Figura 21 - Proceso de la minería de datos.

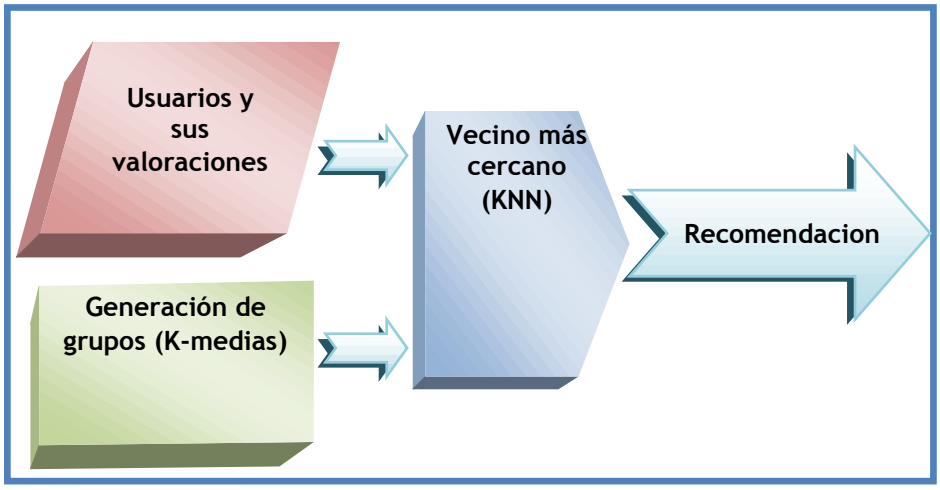


Figura 21 - Proceso de la minería de datos[23]

Nuestros valores de entrada son: los usuarios y las valoraciones de los restaurantes; y, al procesar los valores por medio del algoritmo de k-medias, se generan los grupos de usuarios. Los grupos de usuarios generados siguen un patrón de agrupamiento de tipo "cerca y rellena", que fueron mencionados en la sección 3.4.1 Justificación de la técnica de minería de datos escogida.

Además de los datos y de la tarea, existen otros aspectos que influyen en el aprendizaje y definen nuestros patrones, que suelen denominarse Bias.

- **Bias del lenguaje**, es la manera de expresar o definir los patrones y es el bias que más influye en esta complejidad.

[23] Hernández, José. Introducción a la Minería de Datos. s.l: 2004: Pearson, 2004

Expresamos nuestro modelo de una forma vectorial, donde cada usuario es un vector y cada calificación a un restaurante es una dimensión.

- **Bias de búsqueda**, es el conocimiento previo que puede ayudar a refinar el espacio de búsqueda. Que en nuestro proyecto se encuentra definido por la agrupación de usuarios.

5.3.2 Tareas y métodos

Una de las primeras cosas que debemos tener claro es diferenciar una *"tarea"* de un *"método",* así como destacar las tareas y métodos más relevantes.

Una (un tipo de) tarea de minería de datos es un (tipo de) problema de minería de datos. Por ejemplo, "clasificar las piezas del proveedor X en óptimas, defectuosas reparables y defectuosas irreparables" es una tarea. Concretamente, el tipo de la tarea es clasificación.

Los métodos o técnicas son quienes permiten resolver estas tareas o problema de minería de datos, la tarea de clasificación, por ejemplo, se podría resolver mediante árboles de decisión o redes neuronales, entre otros métodos.

5.3.2.1 Tareas

El sistema "AFI Restaurantes" realizara dos tareas de minería de datos: La clasificación como tarea predictiva y el agrupamiento o clusterización como tarea descriptiva.

- **Tareas Predictivas**: se trata de problemas y tareas en los que hay que predecir uno o más valores para uno o más elementos. Los elementos en la evidencia van acompañados de una salida (clase, categoría o valor numérico) o un orden entre ellos. [24] Dentro de este tipo encontramos la de tarea de clasificación que usaremos en nuestro sistema.
- **Tareas Descriptivas:** En estos casos los elementos se presentan como un conjunto sin etiquetar ni ordenar de ninguna manera. El objetivo por lo tanto no es predecir nuevos datos sino describir a los existentes. [24] Dentro de este tipo de encontramos la tarea de clusterización que usará "AFI Restaurantes".

5.3.2.2 Métodos

Las tareas que realizará el sistema, requiere métodos, técnicas o algoritmos para resolverlas. Una de las cosas que más sorprende es que, además de que, una tarea

[24] Hernández, José. Introducción a la Minería de Datos. s.l: 2004: Pearson, 2004

puede tener muchos métodos diferentes para resolverla, tenemos que él mismo método (o al menos el mismo tipo de técnica) puede resolver un gran abanico de tareas. Esto no es casualidad, sino que se debe a que, en el fondo, la mayoría de las tareas son caras de la misma moneda, *el aprendizaje inductivo*.

Las técnicas de minería de datos, usada por el sistema "AFI Restaurantes", se encuentran dentro de la clasificación de técnicas basadas en casos, en densidad o distancia, que tienen como generalidades las siguientes:

- Se basan en distancias al resto de elementos, como los vecinos más próximos (los casos más similares).
- La técnica es más sofisticada por medio de la estimación de funciones de densidad.
- Se usan algoritmos jerárquicos, como Two-step o COBWEB y los no jerárquicos, como K medias.

La correspondencia que existe entre los métodos que solucionan las distintas tareas posibles en la minería de datos, se muestra detallada en la sección de Anexos. (Ver A.2 Correspondencia entre tareas y métodos)

5.3.3 Minería de datos y aprendizaje inductivo

Hablando de tareas: ¿Qué tiene que ver un agrupamiento con una regresión? y hablando de métodos: ¿Qué tiene que ver un "vecino más cercano" con un árbol de decisión? o ¿Cómo puede ser que todo se encuentre dentro de la "extracción de conocimiento"? Son interrogantes que nos planteamos para definir las relaciones existentes entre las tareas y los métodos que utilizaríamos en "AFI Restaurantes".

La respuesta a estas interrogantes es, reconocer que todas las tareas (exceptuando las reglas de asociación y las correlaciones) y los métodos se centran alrededor de la idea del aprendizaje inductivo y que son presentaciones diferentes del mismo proceso.

¿Qué es el Aprendizaje Inductivo?

El aprendizaje no es sencillo definir ya que es un término que se usa en muchas ciencias y recientemente más en Informática. El aprendizaje inductivo es un tipo especial de aprendizaje que parte de casos particulares (elementos) y obtiene casos generales (reglas o modelos) que generalizan o abstraen la evidencia."[25]

[25] Hernández, José. Introducción a la Minería de Datos. s.l: 2004: Pearson, 2004

5.4 Implantación e impacto de la minería de datos

5.4.1 Introducción

Académicos e investigadores de mercado a menudo encuentran la mejor solución para resolver sus estudios mediante la definición de grupos homogéneos de objetos, ya sean ellos individuos, firmas, productos, o incluso comportamientos. [26] Siguiendo esta línea de pensamiento, consideramos importante definir un objetivo metodológico para el desarrollo del sistema "AFI Restaurantes".

La misma metodología utilizada para resolver las necesidades en otras áreas de estudio como las ciencias físicas para clasificación de varios grupos de animales, o en las ciencias sociales para el análisis de varios perfiles psiquiátricos; puede ser aplicada en sistemas informáticos.

5.4.2 Claves del éxito de un programa de minería de datos

La minería de datos puede ser definida como la "extracción no trivial de información implícita, desconocida previamente, y potencialmente útil a partir de los datos".[27]

---

[26] Enrique Herrera-Viedma & Carlos Porcel & Lorenzo Hidalgo. Sistemas de recomendaciones: Herramientas para el filtrado de información en Internet [en línea], "Hipertext.net", núm. 2, 2004.

[27] Hernández, José. Introducción a la Minería de Datos. s.l: 2004: Pearson, 2004

Las bases de datos actuales acumulan una gran variedad y cantidad de datos, estadísticas e índices en los cuales la información útil no es fácil de encontrar o inferir a simple vista.

"AFI Restaurantes" ha sido desarrollado con fines académicos para la exposición de bondades que ofrece la minería de datos en áreas de predicción y recomendación. Sin embargo, el mismo esquema, que hemos planteamos, podría ser usado por empresas o entidades que están interesadas en rescatar esa información y generar nuevas oportunidades de negocio tales como:

- Mejorar el funcionamiento de la organización.
- Optimizar el manejo de sus bases de datos.
- Predicción automatizada de comportamientos y tendencias.
- Obtener ventajas comerciales.
- Mejorar calidad de productos.
- Descubrimiento automatizado de modelos desconocidos.
- Descubrimiento de anomalías y acciones fraudulentas por parte de clientes.

Las oportunidades son claras y para ello hemos seguido un marco de referencia de "buenas prácticas" o fundamentos para

lograr el éxito de "AFI Restaurantes" como proyecto de minería de datos. Estas buenas prácticas son mencionadas a continuación:

- El negocio[28] y sus necesidades han de dirigir el desarrollo del programa.
- Correcta especificación de problemas concretos y específicos de minería de datos.
- Calidad de datos, ya sea por un programa previo o por limpieza es primordial.
- Uso de herramientas integradas y entornos de usabilidad.
- Equipo heterogéneo de personal capacitado no sólo en minería de datos, sino también en estadística, bases de datos y el campo de acción del negocio.

5.4.3 Formulación del programa: fases e implementación

Como hemos mencionado en la sección anterior, es crucial elaborar un programa que nos permita gestionar el desarrollo de "AFI Restaurantes" como un proyecto de minería de datos. Para dicha formulación, definimos: qué es necesario incluir en el programa, qué estructura se va a seguir y cuál será su extensión y alcances.

[28] Utilizamos la palabra "negocio" para identificar el ámbito y contexto de una organización.

5.4.3.1 El modelo y guía de referencia CRISP-DM[29]

Como modelo y guía de referencia para elaborar el presente proyecto de tesis hemos utilizado la metodología definida en el estándar internacional CRISP-DM, el cual propone un modelo de proceso general para proyectos de minería de datos que es neutral respecto a industria y herramientas, y es aplicable en cualquier sector de negocio.

El estándar 1.0 contiene las respectivas fases de un proyecto, sus respectivas tareas, y las relaciones entre estas tareas.

El ciclo de vida del proyecto consiste de 6 fases. (Ver Figura 22 - Fases del modelo de referencia CRISP-DM) La secuencia de estas fases no es estricta, y regresar o avanzar entre diferentes fases es siempre necesario.

[29] **CRISP-DM**: Estándar de la industria para el proceso de minería de datos respaldado por un consorcio de empresas (inicialmente bajo una subvención inicial de la Comisión Europea) que incluye a SPSS, NCR y DaimlerChrysler. (http://www.crisp-dm.org)

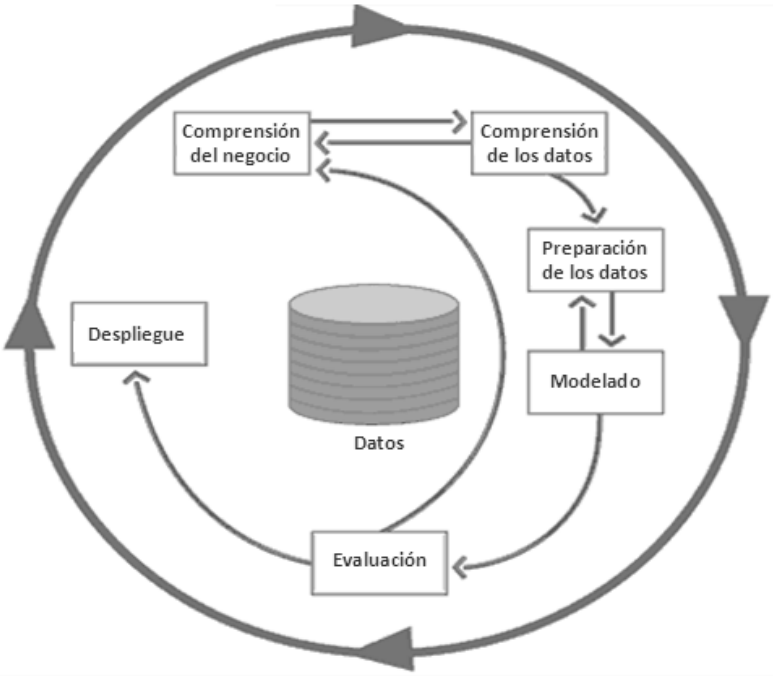


Figura 22 - Fases del modelo de referencia CRISP-DM[30]

A continuación, una breve descripción de lo que fue realizado en cada una de las fases mencionadas:

1. **Comprensión del negocio**

   La fase inicial se concentra en entender el contexto en que se desenvolverá el sistema "AFI Restaurantes" (ver sección 1.3.2 Análisis del usuario y su entorno) y entonces, convertir este conocimiento en una definición de problema de minería de datos (ver sección 1.1 Definición del problema) y una definición

[30] CRISP-DM. Cross Industry Standard Process for Data Mining. [en línea], disponible en: http://www.crisp-dm.org/Process/index.htm

de los objetivos a alcanzar (ver sección 1.2 Objetivos del proyecto).

2. **Comprensión de los datos**

   En la fase de comprensibilidad de los datos definimos el tipo de colección de datos y procedimientos, que utilizaremos en el sistema.

3. **Preparación de los datos**

   La fase de preparación de los datos cubre las actividades de ir construyendo el conjunto de datos, datos que obtienen de manera explícita y alimentan las herramientas de modelado.

4. **Modelado**

   En esta fase, utilizamos las técnicas de k-medias y vecino más cercano para obtener nuestro modelo de predicción y recomendación.

5. **Evaluación**

   En este escenario del proyecto, ya se ha construido un modelo de alta calidad, desde una perspectiva de análisis de datos. Para proceder al resultado final, se evalúa el modelo, y se revisan los pasos a ejecutarse para detectar posibles problemas de rendimiento y darles su respectiva solución. (Ver sección 3.4.2 Justificación de los algoritmos escogidos)

6. **Despliegue**

   El propósito del modelo es incrementar el conocimiento de los datos, el conocimiento ganado necesitara ser organizado y presentado en una manera que el usuario pueda usar y entender[31], esto se cual se logra con el despliegue del sitio Web "AFI Restaurantes".

5.5 Recopilación de datos

Un sistema de recomendaciones se alimenta dinámicamente de la información que recibe de los usuarios que van registrándose en el sistema. Siendo así, la fuente principal de recopilación de información será el sitio web, publicado en un servidor público, donde los usuarios ingresarán la información de interés.

Los datos indispensables que necesitamos recopilar para el entrenamiento de nuestro modelo fueron definidos en los Requerimientos del sistema (ver Capítulo 3.1 Requerimientos del sistema)

Dado que el sistema, en un principio no lo tendríamos montado sobre un servidor de acceso público fue necesario utilizar un método

[31] CRISP-DM. Cross Industry Standard Process for Data Mining. [en línea], disponible en: http://www.crisp-dm.org/Process/index.htm

alternativo para recopilar información para dar inicio al análisis y pruebas. Utilizamos una metodología de encuestas.

La encuesta, cuyo formato puede ser encontrado en la sección de anexos, fue realizada a 25 personas en la que buscaba obtener la siguiente información de los usuarios:

5.5.1 Datos obtenidos por encuestas

- **Nombre de usuario y contraseña**. La contraseña sería una contraseña temporal que podrá ser cambiada una vez el usuario tenga acceso al sitio Web.
- **Dirección de correo electrónico**. Información básica para el registro del usuario. Se usará esta dirección también para notificar al usuario cuando el sistema esté disponible en línea.
- **Edad, Sexo y Factor P**. Esta información fue pedida en caso que se quisiera realizar un análisis en base a dichas variables en un futuro. Para el proyecto actual de tesis, no serán utilizadas.
- **Calificaciones de restaurantes.** Se solicitó al usuario que califique en una escala del 1 al 5 los establecimientos que haya visitado tanto en calidad de servicio y calidad de la comida. Un análisis posterior nos sugirió que no existía una diferencia significativa entre los dos criterios por lo cual en el sistema

hemos considerado el criterio de calificación general del establecimiento.

- **Categoría del restaurante.** Como desarrolladores del sistema, nosotros planteamos una categorización predeterminada para los establecimientos. Sin embargo, los usuarios suelen tener sus propias maneras de categorizar; por lo que hemos considerado adecuado, como retroalimentación, conocer cómo ellos clasifican los restaurantes propuestos.

Actualmente, mediante las encuestas, se ha logrado calificar varios restaurantes de entre los 200 restaurantes disponibles en la base de datos obtenido un aproximado de 500 (quinientas) calificaciones por parte de los usuarios. Este número se logrará incrementar una vez que el sitio Web esté disponible públicamente.

En dicha fase final del programa, los datos se capturan de forma electrónica por medio de un portal Web en donde las personas acceden al sitio y dan sus calificaciones de preferencia, haciendo uso de publicidad en email, pasando la voz entre colegas, amigos y conocidos.

5.6 Representación de un sistema de recomendación

Antes de empezar a explicar los diferentes modelos en que guardaremos la información vamos a aclarar ciertos conceptos.

A partir de ahora, llamaremos usuario a cada uno de los individuos presentes en "AFI Restaurantes" y que desean que les sean realizadas determinadas recomendaciones sobre determinados objetos que formen parte del tema en el que se está basando nuestro sistema de recomendación. Un objeto será por lo tanto un componente más dentro del espacio que estemos tratando.

En un sistema de recomendación colaborativo habrá que representar por una parte los objetos del sistema, y por otra parte tendremos que representar de alguna forma a los usuarios. De esta manera el sistema recomendará al usuario en cuestión, objetos del sistema que no conozca y a los que otros usuarios muy parecidos a él han valorado positivamente.[32]

Consideremos en primer lugar un conjunto O, formado por todos los objetos del sistema:

$$O = \{O_1, O_2, O_3, \dots , O_i, \dots , O_m\}$$

Por otra parte, tendremos otro conjunto U constituido por todos los usuarios:

$$U = \{U_1, U_2, U_3, \dots , U_i, \dots , U_n\}$$

[32] Premios NAI. Memoria SRI, Sistema Inteligente de Recomendaciones (pdf). 2003.

Siguiendo esta representación, cada usuario U_i del sistema podríamos entenderlo como un vector

$$U_i = (Calif_{i1}, Calif_{i2}, Calif_{i3}, \ldots , Calif_{ij}, \ldots , Calif_{in})$$

donde cada componente $Calif_{ij}$ representaría la calificación con la que el usuario U_i ha valorado al objeto O_j.

Un usuario de "AFI Restaurantes" no tiene que haber calificado a todos los objetos del sistema; es más, si esto fuera así, no tendría sentido el sistema de recomendación puesto que no tendría qué recomendar al usuario puesto que ya lo conoce todo.

Por lo tanto, dentro de este vector U_i que representa al usuario, habrá componentes vacías, componentes que el usuario no ha votado debido a que aún no las conoce. Tenemos entonces un espacio vectorial un poco peculiar, con vectores que no tienen valor en todas sus componentes. Este aspecto influirá en la manera en que tendremos que definir las diferentes funciones entre vectores que utilizaremos.

A continuación pondremos un ejemplo, relacionado a nuestro contexto de restaurantes, que aclare los conceptos explicados hasta ahora.

Supongamos que tenemos un sistema de recomendación colaborativo donde el conjunto de objetos a recomendar son los restaurantes de la ciudad de Guayaquil.

Suponiendo que:

- El Usuario1 (U_1) ha frecuentado 6 restaurantes: A, B, D, J, L, V ; y ha calificando respectivamente dichos restaurantes con: 9.5, 3, 3, 8, 7 y 4.
- El Usuario2 (U_2) ha frecuentado por su parte 3 restaurantes: C, J, V; a los que en este caso dio la siguiente valoración: 10, 5 y 6.

Para representar a estos dos usuarios utilizaríamos los siguientes vectores de usuario:

U_1 =(9.5,3,vacio,3,vacio,...,vacio,8,vacio,7,vacio,...,vacio,4, vacio,...,vacio)

U_2 =(vacio,vacio,10,vacio,...,vacio,5,vacio,...,vacio,6, vacio,...,vacio)

De esta manera, vemos que quedarán vacías las coordenadas de un vector que correspondan a libros que desconoce el usuario en cuestión.

Para realizar las recomendaciones a los usuarios en estos sistemas colaborativos representados con un espacio vectorial de este tipo, el sistema elegirá objetos que el usuario desconozca y que tengan una valoración alta en otros usuarios parecidos a él.

Debemos definir entonces métodos que nos sirvan para conocer el grado de similitud entre dos usuarios, ó lo que es lo mismo, grado de similitud entre sus dos vectores representativos. Medidas que pueden servir para este objetivo son el coseno del ángulo formado por ambos vectores, ó la correlación vectorial de ellos.

5.6.1 Modelo abstracto de un sistema de recomendación (SR)

Los SR para generar recomendaciones, usan las entradas del usuario activo, pero también información sobre los ítems o información del resto de usuarios del sistema, que actúan como colaboradores. En este sentido, la realimentación por parte de los usuarios es muy importante de cara a albergar una información más completa ante futuros procesos de generación de recomendaciones. La Figura 23 - Esquema del proceso de generación de una recomendación, refleja el proceso de generación de las recomendaciones.

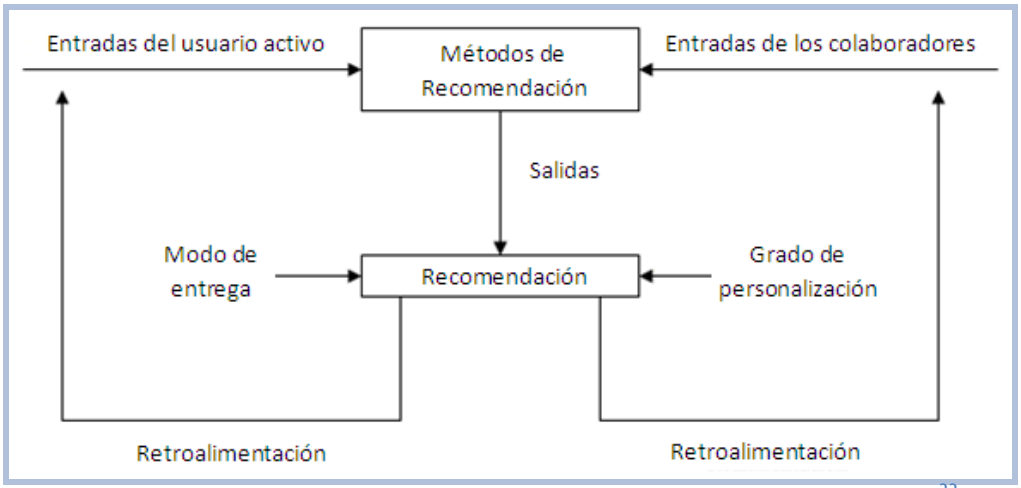


Figura 23 - Esquema del proceso de generación de una recomendación[33]

Para poder realizar una recomendación a un usuario, es necesario conocer algún tipo de información sobre sus preferencias. Además, dependiendo del tipo de sistema también necesitaremos información sobre los ítems a recomendar o información reunida sobre el resto de usuarios del sistema (comunidad de usuarios o colaboradores). Esta información que necesitamos para realizar las recomendaciones constituye la entrada o entradas del sistema.

La información sobre los usuarios puede venir dada de dos formas, sin necesidad de ser mutuamente exclusivas: por extensión o intencionalmente. Por extensión se refiere a información que se tenga sobre las experiencias pasadas del

[33] Enrique Herrera-Viedma & Carlos Porcel & Lorenzo Hidalgo. Sistemas de recomendaciones: herramientas para el filtrado de información en Internet [en línea]. "Hipertext.net", núm. 2, 2004.

usuario con respecto a los ítems encontrados. Es lo que también conocemos como navegación implícita pues el usuario no es consciente de estos seguimientos. Por información expresada intencionalmente se entiende alguna especificación de los ítems deseados por los usuarios. También se llama navegación explícita y consiste en que el usuario expresa intencionalmente (de forma explícita) al SR información sobre sus preferencias.

La salida del sistema está constituida por las recomendaciones generadas por el sistema, que variarán dependiendo del tipo, cantidad y formato de la información proporcionada al usuario.

Algunas de las formas más comunes de representar la salida son las siguientes:

- Sugerencia o lista de sugerencias al usuario de una serie de ítems.
- Presentar al usuario predicciones del grado de satisfacción que se asignará al ítem concreto. (Estas estimaciones pueden ser presentadas como personalizadas al usuario o como estimaciones generales del conjunto de colaboradores).
- Cuando la comunidad de usuarios es pequeña o se conocen bien los miembros de dicha comunidad, podría

ser útil visualizar las valoraciones individuales de los miembros que permitiría al usuario activo obtener sus propias conclusiones sobre la efectividad de una recomendación.

- Independientemente de estos formatos de salida, puede resultar muy interesante incluir una breve descripción o explicación sobre el ítem recomendado a modo de justificación del porqué de dicha recomendación.

5.6.2 Representación de nuestro sistema de recomendación

Una vez expuesta de manera abstracta, la forma en que podemos representar los componentes de un sistema de recomendación, a continuación mostramos la traslación de este modelo a nuestra aplicación Web.

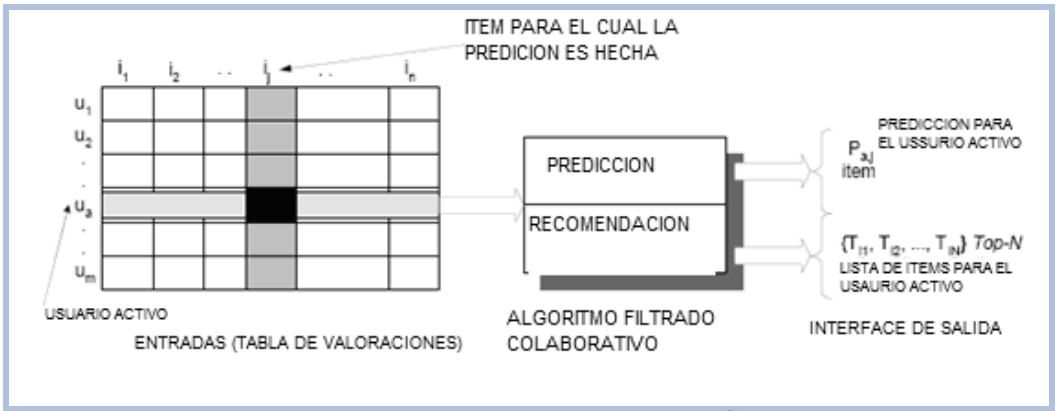


Figura 24 - Proceso del filtrado colaborativo[34]

[34] Sarwar, Karypis, Konstan, Riedl. GroupLens Research Group/Army HPC Research Center, ItemBased Collaborative Filtering Recommendation Algorithms.

La información que representa a los objetos y a los usuarios de nuestro sistema de recomendación reside en la base de datos utilizada para ese fin.

Los objetos que componen nuestro sistema existen como registros de una tabla de la base de datos que denominamos Restaurante. Por otra parte, los datos correspondientes a los usuarios (nombre, etc.) constituyen los registros de otra tabla de la base de datos, la tabla Usuario[35]. Como señalamos en líneas anteriores, un Usuario del sistema debe representarse como un vector formado por las votaciones de los restaurantes que haya votado, teniendo vacías estos vectores aquellas componentes que representen a los objetos que no haya votado. (Ver Figura 25 - Ejemplo de representación de nuestro modelo)

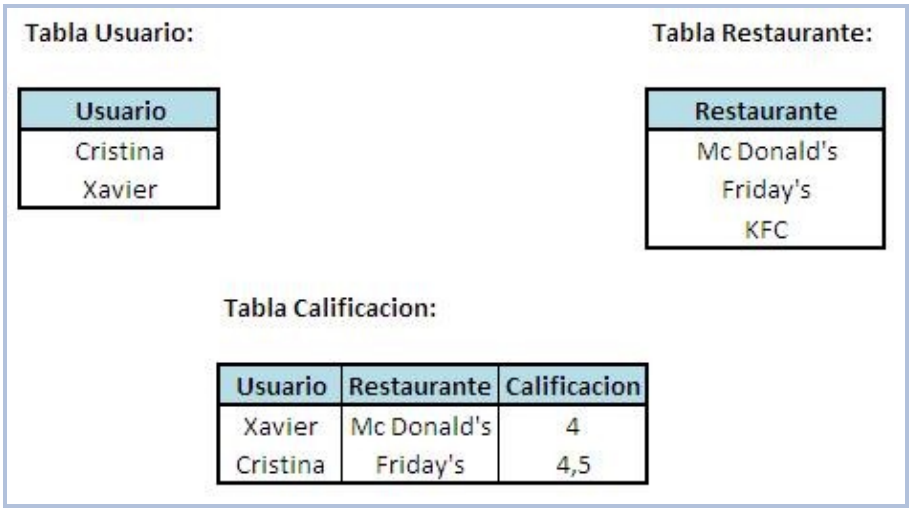


Figura 25 - Ejemplo de representación de nuestro modelo

[35] Usaremos para el ejemplo una representación abstracta de la tabla de usuarios real.

Para representar estos vectores de usuarios (Ver Figura 25 - Ejemplo de representación de nuestro modelo), existe una tabla más en la base de datos del sistema, denominada Calificación que estable una relación directa Usuario-Restaurante, en la que cada registro representará una votación de un objeto concreto por un individuo.

Con esta representación, los vectores teóricos correspondientes a los usuarios Xavier y Cristina serían:

Xavier = (4, vacio, vacio)

Cristina = (vacio, 4.5, vacio)

Los algoritmos utilizados en nuestro sistema para calcular el grado de similitud entre los diferentes usuarios, así como aquellos algoritmos que calculan los pronósticos y las recomendaciones extraen las componentes de dichos vectores a partir de los datos de estas tablas de la base de datos.

El mecanismo en el que se basa nuestro sistema para realizar recomendaciones se basa en calcular qué usuarios son los más parecidos a cada uno, y a partir de ahí se recomendarán aquellos objetos que obtuvieron una calificación alta en estos usuarios y que el usuario en cuestión no votó.

Para generar el grado de similitud entre dos usuarios, se utilizarán el algoritmo de KNN, en cuyo procedimiento será

necesario usar la medida del coseno del ángulo entre vectores representativos. A continuación mostramos cómo se calculamos estas medidas entre los vectores de usuarios:[36]

1. **Coseno del ángulo:** es una medida utilizada para medir el grado de similitud entre dos vectores. A mayor valor de esta medida significará que más parecidos son. Recordamos que el valor mínimo de esta medida será "-1" y el valor máximo "1". La manera de calcular el coseno del ángulo sería:

   $$\cos\theta = \frac{(x.y)}{(|x||y|)}$$

   Donde $(x.y)$ sería el producto escalar entre los 2 vectores, y $|x||y|$ sería el producto de sus módulos. Debido a que nuestro espacio vectorial es algo peculiar (puesto que tenemos componentes en los vectores sin valor en algunos de ellos) sólo consideraremos para cualquier cálculo las componentes de los vectores para las que existe valor en ambos.

[36] Premios NAI. Memoria SRI, Sistema Inteligente de Recomendaciones (pdf). 2003.

5.7 Proceso para el cálculo de pronósticos

Uno de los servicios proporcionados por nuestra aplicación es el cálculo de pronósticos a un usuario sobre objetos del sistema que no conozca.

El proceso que seguimos para realizar estos pronósticos es el siguiente:[37]

1. Un usuario A está identificado en el sistema y le corresponde un vector en memoria de dimensión *n*, donde *n* es el número de restaurantes registrados en la base de datos, cuyos valores son las votaciones de los restaurantes. Estos componentes podrán tener un valor (en el caso de que las haya votado) ó estar vacías (en el caso de que no las haya votado).
2. El sistema predice mediante el algoritmo de Fisher a qué grupo de vectores el usuario A pertenece, grupos que fueron generados anteriormente usando el algoritmo de k-medias. El número de grupos existentes en el sistema es un valor que se obtiene usando la prueba F de reducción de variabilidad. Cada grupo estará representado por un vector de medias denominado centroide.
3. El algoritmo del vecino más cercano obtiene una lista ordenada por el grado de similitud entre vectores de usuarios usando como criterio el coseno del ángulo.

[37] Premios NAI. Memoria SRI, Sistema Inteligente de Recomendaciones (pdf). 2003.

4. Los registros de la lista hacen referencia a cada usuario del grupo y contiene la valoración hecha de cada restaurante. De esta lista, se extrae un top N de aquellos restaurantes que los usuarios del grupo han calificado pero el usuario A no lo ha hecho.
5. Debido a que los usuarios van realizando votaciones durante la vida del sistema, es claro que sus vectores representativos no pertenecerán siempre a los mismos grupos. Periódicamente el administrador del sistema pone en funcionamiento el algoritmo de clusterización (k-medias) que agrupa por clases a todos los usuarios.

CAPÍTULO 6

Análisis financiero

6.1. Análisis comercial

"AFI Restaurantes" ayudan al usuario a seleccionar elementos de una gran cantidad de opciones. Una de las variables importantes a considerar para su aceptación comercial es la efectiva gestión de grandes volúmenes de información, ya que esto determina el detalle y calidad de las recomendaciones. Factores como el tiempo de vida (del elemento a evaluar), el tipo de elemento (películas, gente, artículos, etcétera) y la cantidad generada influyen de manera directa en el momento de la recomendación.

Existen también implicaciones sociales. Establecer un perfil de intereses de las recomendaciones puede ocasionar problemas en casos de imparcialidad. Además, la privacidad de los participantes debe considerarse ya que no todos los usuarios gustan de exhibir sus preferencias o de no ser reconocidos por su aportación. Debido a que

mantener un sistema de recomendación es caro, se han considerado diferentes modelos para costear dichos sistemas:

- El consumidor paga por el servicio.
- Los anuncios de publicidad mantienen el sistema.
- El dueño del elemento a evaluar paga por la evaluación de su elemento.

6.2. Análisis de costos

Para el desarrollo e implementación del sistema de recomendación y predicción "AFI Restaurantes" se decidió optar por el uso de herramientas de evaluación que teníamos disponibles. Cabe recalcar, que se trata de software versión *trial*es originales. Usaremos las que listamos a continuación:

1. **Motor de Base de Datos:** Microsoft SQL Server 2005 Express Edition.
2. **Plataforma de Desarrollo:** Visual Studio 2005 Profesional Edition (versión de evaluación de 90 días)
3. **Diseño:** Visio 2003 Profesional y Poseidon UML

Dado que la aplicación debería funcionar en Internet, una vez que sea puesta en producción, debemos analizar costos de la infraestructura de soporte para el sitio web.

Los costos dependerán de la tecnología en la cual las páginas hayan sido diseñadas (páginas .aspx en nuestro caso) y también del motor de base de datos sobre el cual estaremos montando nuestro sistema de recomendación y predicción.

En la fase de desarrollo, el motor de base de datos SQL Server 2005 Express Edition que hemos utilizado, cumple efectivamente su labor para la implementación.

Sin embargo, debemos considerar que en la fase de producción el motor de base de datos deberá soportar una mayor carga transaccional, de manera que se debería considerar la actualización a un motor tipo servidor, con mejores características de rendimiento, seguridad y respaldo. Si se trata de un motor de base de datos que trabaje bajo licencia, esto representaría un costo adicional el cual no ha sido contemplado en la presente tesis de grado. Supondremos el uso de un RDBMS[38] gratuito como mySQL para la fase de producción.

Dirigiéndonos directamente a los costos, nuestro costo más significativo será el alojamiento Web que bordea, en promedio, los 25 dólares anuales. La ventaja de utilizar este tipo de servicio es que no incurriremos en gastos de mantenimiento de equipos, mantenimiento y capacitación de personal de soporte, uso de tiempo

[38] RDBMS: (inglés) Sistema de Administración de Base de Datos Relacional

y espacio físico. El proveedor de alojamiento Web cubre dentro de su paquete de servicio todo lo anteriormente mencionado.

Se considera que existirían costos adicionales en caso que algún interesado, en particular, quisiera implementar el concepto de sistemas de recomendación que planteamos en otro tipo de productos de su interés; para lo cual podría migrar hacia otras plataformas de base de datos como Oracle, Informix, DB2; y otro tipo de páginas Web como PHP, JSP, entre otras.

6.2.1. Costo de inversión

La inversión de nuestra aplicación para fines demostrativos no incide en mayores costos que los siguientes:

Concepto	Tipo	Valor
SQL Server 2005 Express Edition	Gratuito	$ -
Visual Studio 2005 Profesional	Evaluación	$ -
Hospedaje Web*	Contratado	$ 25,00
Total		$ 25,00

* El proveedor de servicios de hospedaje Web da soporte para el motor de base de datos necesario.

Tabla 9 - Costos de inversión

6.2.2. Costos de licenciamiento

Considerando el caso de una implementación independiente, donde una infraestructura debe ser montada por completo, se incurriría en costos de licenciamiento tanto para el servidor

base, el motor de base de datos y la plataforma de desarrollo. Para nuestro caso de implementación de AFI Restaurantes estos costos han sido cero por ser versiones de evaluación y/o gratuitas.

Concepto	Tipo	Valor
Equipo Servidor	Comprado	$ 800.00
Windows Server 2003 Web Edition	Licenciado	$ 399.00[39]
SQL Server 2005 Workgroup Edition	Licenciado	$ 3700.00[40]
Visual Studio .NET 2005 Standard Edition	Licenciado	$ 299.00[41]
Dominio Web	Alquilado(anual)	$ 20,00
Total		**$ 5218,00**

Tabla 10 - Costo de inversión (plataforma independiente)

La Tabla 10 - Costo de inversión (plataforma independiente), muestra un escenario íntegro de soluciones Microsoft para la ejecución comercial del sistema. Sin embargo, la diversidad de plataformas de desarrollo de aplicaciones web y motores de base de datos disponibles es tan amplio que queda a libre elección del interesado dependiendo de la robustez, velocidad y/o poder de procesamiento que satisfaga los objetivos de su implementación.

[39] Microsoft Corporation. Microsoft Windows Server 2003 R2: How to buy [en línea], http://www.microsoft.com/windowsserver2003/howtobuy/licensing/pricing.mspx
[40] Microsoft Corporation. Microsoft SQL Server: How to buy [en línea], http://www.microsoft.com/sql/howtobuy/default.mspx
[41] Microsoft Corporation. MSDN Visual Studio: How to buy [en línea], http://msdn.microsoft.com/vstudio/howtobuy/default.aspx

6.2.2. Costos operativos

Existen varios costos operativos que quizás pasen desapercibidos pero igualmente necesarios y que influyen enormemente en costo final de la aplicación y si no se manejan los tiempos y compromisos adecuados de desarrollo y depuración el costo final puede no ser el esperado:

Las tablas muestran los costos en que se incurrieron durante la etapa de desarrollo de la aplicación de nuestra tesis (aproximadamente de 5 meses), los valores estipulan el valor promedio realizados por mes en los diferentes gastos como gastos de comunicación, de trasporte, y desarrollo. 8

Los gastos de comunicación se refieren al incremento en las facturas propias de cada integrante en los servicios que cada uno contrata para mantener la comunicación. Esto servicios fueron: el acceso a Internet y el uso de telefonía fija y celular para establecer las reuniones, citas y presentaciones; así como para la investigación del tema de nuestro proyecto de tesis (artículos, programas, código fuente, algoritmos).

Gastos de comunicación	Integrante 1	Integrante 2	Integrante 3	Total parcial
Servicio Internet (banda ancha y dial-up)	$ 60.00	$ 60.00	$ 20.00	$ 140.00
Telefonía fija	$ 25.00	$ 25.00	$ 25.00	$ 75.00
Telefonía celular	$ 10.00	$ 10.00	$ 10.00	$ 30.00
Total				$ 245.00

Tabla 11 - Gastos de comunicación (mensual promedio)

Los gastos de transporte se refieren al gasto obligatorio que cada integrante del equipo tuvo que realizar para las reuniones en un lugar especifico; en este caso, se escogió el domicilio de unos de los integrantes por ser estratégico para las 3 personas. Cada miembro incidía en este gasto ya sea usando el servicio de taxi, bus o vehículo particular. Este gasto considera la acción de trasladarse de su lugar de trabajo, y domicilio al lugar de reunión así como los traslados para hacer presentar los avances del proyecto de tesis.

Gastos de transporte	Integrante 1	Integrante 2	Integrante 3	Total parcial
Taxi	$ 0.00	$ 0.00	$ 10.00	$ 10.00
Bus	$ 0.00	$ 0.00	$ 4.00	$ 4.00
Vehículo particular (gasolina)	$ 10.00	$ 10.00	$ 0.00	$ 20.00
Total				$ 34.00

***Asumiendo 3 días promedio de reunión más 2 días de presentación avances**

Tabla 12 - Gastos de transporte (mensual promedio)

Los gastos de desarrollo hacen referencia al gasto incurrido en consumibles como: resmas de papel, cartucho de tóner de impresora, plumas y demás elementos necesarios en el análisis, diseño y presentación del proyecto. Hubo también gastos de alimentación por almuerzos, fuera de nuestros hogares, en sitios de comida rápida y restaurantes. Se incurrió en un incremento de consumo de energía eléctrica que cada uno experimento por concepto de mantener los equipos de computo encendidos más horas de las normales; así como por el uso de acondicionador de aire, impresoras, celulares (cargadores) en un promedio de 6 horas diarias cuando se daban las reuniones de desarrollo, tiempo de uso del computador para investigación, etc.

Gastos de desarrollo	Integrante 1	Integrante 2	Integrante 3	Total parcial
Alimentación*	$ 0.00	$ 0.00	$ 10.00	$ 10.00
Energía eléctrica** (equipos, A/C, impresoras)	$ 25.00	$ 40.00	$ 15.00	$ 80.00
Consumibles (papel, cartucho de tóner, plumas, lápices, entre otros)	$ 10.00	$ 10.00	$ 10.00	$ 30.00
Total				$ 120.00

*** Asumiendo 4 días promedio de reunión.**
**** El gasto eléctrico es el recargo extra en sus consumos por usar cada integrante sus equipos en sus aposentos**

Tabla 13 - Gastos de desarrollo (mensual promedio)

6.3. Productos similares existentes en el mercado

En esta sección vamos a analizar algunos ejemplos de Sistemas de Recomendación existentes en el medio que han servido de base para numerosos estudios:

- **PHOAKS [42] :** Se trata de un sistema experimental para solucionar el problema de encontrar información relevante y de alta calidad en la Web, usando el enfoque colaborativo en el que los usuarios recomiendan determinados ítems a otros usuarios. PHOAKS trabaja reconociendo, concordando y redistribuyendo automáticamente recomendaciones de recursos Web extraídos de mensajes de noticias.
- **Referral Web[43]:** Numerosos estudios muestran que una de las formas más efectivas de divulgar información y conocimientos dentro de una determinada organización es a través de una red informal de colaboradores o amigos. Referral Web se basa en la idea de combinar 'redes sociales' con el filtrado colaborativo, entendiendo por 'redes sociales' grupos de personas vinculadas por determinadas actividades profesionales.

[42] Phoaks: People Helping One Another Know Stuff [en línea]
disponible en: http://www.cs.indiana.edu/~sithakur/l542_p3/index.html
[43] ReferralWeb: Combining Social Networks and Collaborative Filtering [en línea], disponible en: http://citeseer.ist.psu.edu/kautz97referralweb.html

- **FAB**[44]**:** Sistema orientado a la recomendación de URL que combina el uso de información por extensión con el enfoque colaborativo.
- **CiteSeer**[45]**:** Recomienda páginas Web relevantes y usa las listas de favoritos y la organización de registros, como una declaración implícita de intereses respecto al contenido subyacente y se va midiendo el grado de solapamiento con las de otros usuarios.
- **GroupLens**[46]**:** El proyecto GroupLens diseña, implementa y evalúa un sistema de filtrado colaborativo para Usenet, un servicio de listas de discusión con un alto volumen de negocio en Internet.

[44] Balabanovic, FAB: An Adaptive Web Page Recommendation Service [en línea], disponible en: http://dbpubs.stanford.edu:8090/pub/2000-40
[45] CiteSeer.IST: Scientific Literature Digital Library [en línea], disponible en: http://citeseer.ist.psu.edu/cs]
[46] GroupLens Home Page [en línea], disponible en: http://www.grouplens.org/

Conclusiones y recomendaciones

El proyecto de “AFI Restaurantes” empezó como una idea planteada por uno de los integrantes del grupo quien había escucha de sistemas de predicción y recomendaciones durante los cursos de Ingeniería de Software dictados en la FIEC.

El tema cobró verdadero interés durante el tópico de graduación de “Aplicaciones de minería de datos” dirigido por el Msc. Fabricio Echeverría, donde se empezó a trabajar la idea para ser presentada como tesis de grado.

El proyecto empezó siendo algo pequeño, sin embargo, nuevas ideas y profundidad en los temas se fueron mostrando conforme avanzó el desarrollo del tópico. El desarrollo de las etapas de análisis y diseño del sistema fue guiado por el director del tópico y revisadas, cuidadosamente, por los integrantes del grupo; presentándose

continuamente, hasta el final, cambios y sugerencias en el contenido y diagramación.

Un punto importante a considerar fue la necesidad de investigar soluciones para resolver los problemas de escalabilidad a los que está sujeto este tipo de sistemas; aspecto que no era de nuestro conocimiento al momento de embarcarnos en el desarrollo de este tema de tesis.

Durante la implementación de nuestro sistema "AFI Restaurantes" pudimos notar que existe un amplio espectro de aplicaciones vinculado a la minería de datos y, particularmente relacionadas al análisis de conglomerados o clústeres. Con el uso correcto de estas herramientas, hemos descubierto un potencial con el que antes solo podíamos imaginar, la fusión del poder de procesamiento de estos días en conjunto, con el poder de análisis matemático de la Estadística, permiten obtener actualmente información valiosa e interesante para pre y post análisis; lo que antes sólo era un historial o repositorio de datos, se vuelve ahora información de relevancia que ayuda a tomar decisiones.

Recomendamos siempre centrarse en un buen diseño o perspectiva al atacar algún problema en particular, siempre remitiéndose a la teoría

o mejores prácticas para obtener resultados esperados o acordes en tiempos esperados o adecuados. El modelo y guía de referencia CRISP-CM es recomendado como estándar para la planeación, organización, dirección y control de un proyecto de minería de datos.

Recomendamos que las herramientas de desarrollo y algoritmos que se utilicen, tengan optimizadas las operaciones de minería de datos; en la misma línea recomendamos el uso de paralelismo o hilos (*threads*) en los procesos del análisis de conglomerados, proporcionalmente al tamaño de su repositorio de datos, para mantener el rendimiento y desempeño de sus aplicaciones desarrolladas.

Bibliografía

Hernández, José. Introducción a la Minería de Datos. s.l: Pearson, 2004

Peña, Daniel. . Análisis de Datos Multivariantes. s.l: McGraw Hill, 2002

Tercera Edición del Premio NAI. . Memoria SRI, Sistema Inteligente de Recomendaciones (pdf). 2003.

Sarwar, Karypis, Konstan, Riedl. . Item Based Collaborative Filtering Recommendation Algorithms. GroupLens Research Group/Army HPC Research Center, 2003

Referencias bibliográficas

[C2_4] Goldberg, D, Oki, B., Nichols, D., & Terry, D.B. (1992) Using Collaborative Filtering to Wave an Information Tapestry. Communications of the ACM, Diciembre 1992. Vol. 35, nº 12, pp. 61-70.

[C2_5] Resnick, P., Lacovou, N., Sushak, M., Bergstrom, P., & Riedl, J. (1994).GroupLens: An open architecture for collaborative filtering of netnews. Proceedings of the 1994 Computer Supported Collaborative Work Conference.

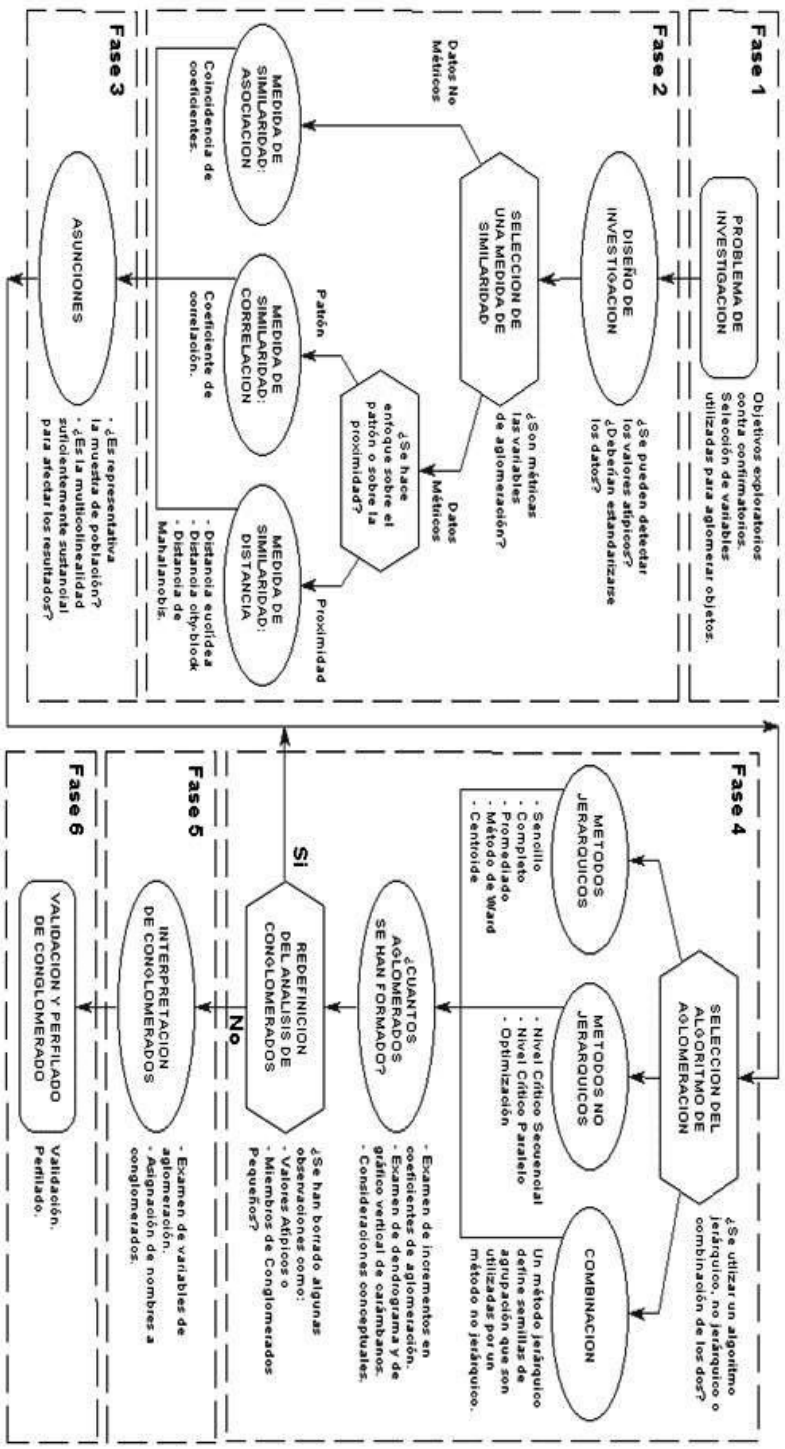
[C2_6] Shardanand, U. & Maes, P. (1995). Social Information Filtering: Algorithms for Automating "Word of Mouth". Proceedings of CHI'95, Denver, Colorado, USA. Mayo, 7-11, 1995. pp. 210- 217.

[C2_7] EachMovie. Recommendation system [en línea], http://www.research.compaq.com/SRC/eachmovie.

[C2_8] Amazon.com. Our site Features> Recommendations [en línea], http://www.amazon.com

Anexos

A.1 Pasos del análisis de clúster o conglomerados



A.2 Correspondencia entre tareas y métodos

Nombre	Predictivo		Descriptivo		
	Clasificación	Regresión	Agrupamiento	Reglas de Asociación	Correlaciones / Factorizaciones
Redes neuronales	✓	✓	✓		
Árboles de decición ID3, C4.5, C5.0	✓				
Árboles de decisión CART	✓	✓		✓	
Otros árboles de decisión	✓	✓	✓	✓	
Redes de Kohonen			✓		
Regresión lineal y logarítmica		✓			✓
Regresión logística	✓			✓	
K-Means			✓		
A priori				✓	
Naive Bayes	✓				
Vecinos más proximos	✓	✓	✓		
Análisis factorial y de comp. Principales					✓
TwoStep, Cobweb			✓		
Algoritmos genéticos y evolutivos	✓	✓	✓	✓	✓
Máquinas de vectores soporte	✓	✓	✓		
CN2 rules (cobertura)	✓			✓	
Análisis discriminante multivariante	✓				

A.3 Diseño lógico de la base de datos