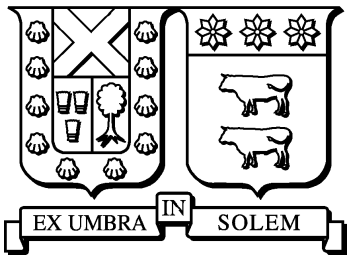
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INFORMÁTICA

SANTIAGO – CHILE



"SISTEMA RECOMENDADOR BASADO EN CONFIANZA PARA CENTROS CULTURALES EN LA REGIÓN METROPOLITANA"

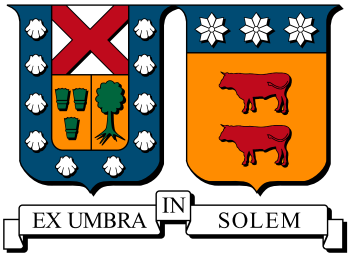
SEBASTIÁN ANDRÉS VIDAL MORENO

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: HERNÁN ASTUDILLO

JUNIO 2016

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO – CHILE



"SISTEMA RECOMENDADOR BASADO EN CONFIANZA PARA CENTROS CULTURALES EN LA REGIÓN METROPOLITANA"

SEBASTIÁN ANDRÉS VIDAL MORENO

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: HERNÁN ASTUDILLO

PROFESOR CORREFERENTE: PABLO CRUZ

JUNIO 2016

Agradecimientos

Resumen

Abstract

Índice de Contenidos

Índice de cuadros

Índice de figuras

Glosario

Introducción

0.1. Contexto

Capítulo 1

Definición del Problema

Actualmente, la colaboración, interacción e intercambio de información son la fuerza que mueve las aplicaciones web de hoy en día (A esto se le conoce como Web 2.0). Ejemplos de esto son los blogs, redes sociales, wikis y sitios específicamente desarrollados para compartir contenido creado por usuarios, tales como Youtube, Deviantart , Twitch o Newgrounds. Con el auge de los dispositivos móviles se ha mantenido esta tendencia a Ya que ahora, en gran volumen, los usuarios cumplen roles de consumidores y creadores, se hace necesario un mecanismo que permita diferenciar a los usuarios y el contenido que producen, para facilitar el enlace entre un usuario y el contenido que éste desea. Con esta motivación, nace el área de sistemas de recomendación basados en confianza, sustentados en áreas más antiguas, como los sistemas de recomendación y reputación. Ya que cualquier persona puede decir cualquier cosa sobre cualquier cosa, uno de los problemas que se presentan, tanto para un usuario común, como para los sitios web, es que dado el alto volumen de información generada, es difícil determinar su nivel de veracidad (ni las intenciones del usuario que las genera).

Capítulo 2

Estado del Arte

2.1. Estado del Arte

La colaboración y la interacción entre usuarios de la web han sido claves para el cambio de paradigma que ocurrió para el uso de Internet a principios del siglo XXI, después del estallido de la burbuja "punto com" a finales del año 2001. Hasta ese entonces, el uso de Internet consistía en navegación de los contenidos de una página web estática, para luego pasar a la siguiente. El paradigma cambia, de sólo lectura a lectura y escritura, de compañías a comunidades y de portales a plataformas web.[**?**]

Entre los sitios que sobrevivieron a la crisis están gigantes como Amazon, Google e eBay. Lo que estas empresas tienen en común es que motivaron una participación activa de sus usuarios, generando un ciclo de retroalimentación que aumentó el número de usuarios,

Con el tiempo, dado el alto volumen de información generada, surgen varios problemas que impiden un uso óptimo de los recursos que ofrece la web. Entre ellos, se pueden encontrar los siguientes, que serán el énfasis en esta memoria.

- En plataformas web, un usuario puede ser abrumado por el exceso de contenidos.
- Las fuentes de información son numerosas y, muchas veces, contradictorias.

En respuesta a estas necesidades, existen sistemas capaces de navegar por el contenido de una plataforma web y hacer recomendaciones de ítems a un usuario, segun los datos disponibles de éste. Además se han desarrollado modelos que describen la confianza entre usuarios de un plataforma, con el objetivo de discriminar entre las diversas fuentes y determinar la más creíble para una persona en particular.

A continuación se detallarán las implementaciones de ambas soluciones y las formas en que ambas pueden enriquecerse mutuamente.

2.1.1. Computación Basada en Confianza

La confianza es un fenómeno social que se da comúnmente en una comunidad de personas, segun [**?**], se dice que el ser humano no podria enfrentar las complejidades del mundo sin recurrir a la confianza, ya que es con confianza que somo capaces de razonar de forma sensata a los eventos de la vida diaria. Por ejemplo, cada día salimos de nuestros hogares *confiando* en que seremos capaces de volver y *confiando* en que no tendremos un accidente que nos lleve al hospital.

En términos generales, existen varias definiciones del concepto de çonfianza", según Morton Deutsch en 1962 el comportamiento de confianza ocurre cuando un individuo percibe un camino ambiguo, el resultado de cual puede ser beneficioso o dañino, y la ocurrencia de un resultado "bueno.º "malo"depende de las acciones de otra persona. Si el individuo elige ir en ese camino, se puede decir que actuó de manera confiada, de lo contrario es desconfiado. Esta definición obedece a la noción de que la estructura básica de la confianza consiste en una decisión.

Otra definición de çonfianza", por Gambetta sugiere que "...confianza, (o simétricamente, desconfianza) es un nivel particular de probabilidad subjetiva de que un agente realice una acción en particular, antes de que pueda monitorear dicha acción y en una situación que afecte su propia acción"[**?**] . En esta defición es importante aclarar que la probabilidad subjetiva es aquella en que un evento es asignado por el individuo basandose en la evidencia disponible (el individuo asigna la probabilidad en base a su experiencia).

Por último, otra definición que se rescata en la literatura está en [**?**], la cual es como sigue ”La confianza es un estado psicológico que involucra expectativas sobre la competencia, benevolencia, integridad y predictibilidad de otra persona y su voluntad de actuar en base a esas expectativas. Los problemas de confianza surgen en contextos que involucran riesgo, vulnerabilidad, incertidumbre e interdependencia. Las expectativas de confianza se crean principalmente por interacciones de las cualidades percibidas de en quien se confía y factores contextuales que entran en juego cuando se hacen decisiones de confianza”.

Lo que se puede destacar de éstas definiciones es la naturaleza de la confianza, que consiste en la toma de decisiones en contextos de incertidumbre y riesgo, basado en valores arbitrarios. Y es una manera en que comprender y adaptarse a las complejidades del medio en que nos desenvolvemos. Siguiendo con [**?**], se desataca que la confianza es lo que juega un rol importante en la formación de grupos, su estructura y los comportamientos que se dan dentro de éstos.

En general se identifican tres tipos de confianza, disposicional, impersonal e interpersonal.

- La confianza disposicional se refiere a un estado interno del individuo, que está dispuesto a confiar. Esta confianza es abierta e independiente de factores externos u otro contexto. Este tipo de confianza se puede subdividir en dos tipos, el tipo A se refere a la creencia del çonfiador.[en] la benevolencia del otro, mientras que el tipo B es la ”disposición que, independiente de la benevolencia del otro, se puede obtener un resultado más positivo actuando *como si* se confiase en el otro. [**?**]

- La confianza interpersonal se refiere a aquella en propiedades percibidas o fe en el sistema o institución donde existe la confianza. Un ejemplo de esto es el sistema monetario [**?**].

- Finalmente, la confianza interpersonal se refiere a la que tiene una persona sobre otra directamente, ésto se puede ver también como una confianza disposicional hacia un sistema viviente. Este tipo de confianza depende del contexto y del individuo. Una persona A puede confiar en una persona B para que arregle su computador, pero no para consejos emocionales.

Como se mencionó anteriormente, la confianza no es una propiedad objetiva, sino un grado subjetivo de creencia en una persona, proceso u objeto. El grado puede variar desde confianza ciega a completa desconfianza, o incluso se puede no tener una opinión sobre el nivel de confianza en otra persona (por falta de información). Este nivel es variable en el tiempo, ya que la confianza en otra persona puede aumentar o disminuir según las experiencias que hayamos tenido con ésta.

Otro aspecto importante de la confianza es que puede aumentar y reforzarse con el uso y disminuir con el desuso, esto se relaciona con la forma en que se razona el proceso de búsqueda de evidencias para confiar o desconfiar, siendo los datos más recientes los más vigentes y confiables. [**?**]

En el campo de la informática, se conoce al modelamiento de la confianza entre personas como Trust-Based Computing[1] [**?**]. En la literatura [**? ? ?**] se menciona que uno de los modelos más antiguos de confianza fue propuesto en el año 1994, por Stephen Paul Marsh, de la Universidad de Strirling [CITA]. Este modelo está establecido sólo con interacciones directas en el proceso de confianza.

Se separa la confianza en tres aspectos distintos, el básico, general y situacional.

- La confianza básica modela la confianza disposicional, mencionada anteriormente. Se calcula a partir de todas las experiencias acumuladas del agente. Es decir que experiencias buenas llevan a mayor confianza y viceversa. Se usa la notación T_x^t para denotar la disposición de un agente x en un momento t
- La confianza general es la que tiene un agente sobre otro, sin tomar en cuenta alguna situación en específico. Representa simplemente la confianza general en otro agente. Se denota $T_x(y)^t$, la confianza general de un agente x sobre un agente y en un tiempo t
- La confianza situacional es la que un agente tiene sobre otro teniendo en cuenta una situación específica. Se calcula considerando la utilidad de la situación, su importancia y la confianza *general*

[1] Computación Basada en Confianza, del inglés

Sin embargo, este modelo tenía algunas limitaciones [**?**]

En primer lugar, consideraba a la persona en quien se confia como una entidad pasiva. Toda acción es de parte de el *confiador*. Es él o ella quien toma la decisión confiar, después de una serie de cálculos y otros factores . Sin embargo, en el mundo real, la persona a confiar nunca es pasiva, ya que envía señales constantemente al confiador, tanto positivas como negativas.

La otra limitación que tenía este modelo consiste en que no se captura el rol de la situación en que se confia. En [**?**] se explica esto con un ejemplo. Una persona A considera entrar a un taxi de B, quien viene saliendo de un bar. Si B ha llevado a A de forma segura en el pasado numerosas veces antes, y B no muestra señales de estar borracho.¿Confiaría A en B de la misma forma? Lo más probable es que no, incluso aunque B cumpla todos los criterios que exige el modelo.

Tampoco toma en cuenta el rol importante que juega la comunicación en la determinación de confianza.

Los modelos de confianza que se utilizan actualmente son varios y se pueden clasificar de muchas formas. Una de ellas está descrita en [**?**], donde se ordenan los modelos según su acercamiento, que puede ser probabilístico o gradual.

- Los modelos probabilísticos operan a un nivel ”todo o nada”, un agente o fuente puede ser confiable *o no*. Muchas veces se basa este tipo de confianza en el número de transaccones positivas o negativas. Estos modelos se utilizan en redes P2P.
- Un modelo gradual hace una estimación de valores de confianza, cuando el resultado de una acción puede ser positiva hasta cierto nivel, en vez de ser *bueno* o *malo*. Esto representa mejor la forma en que los seres humanos interpretan la confianza. Una persona no opera en términos de *confío* o *desconfío*, sino mas bien se confía *mucho*, *mas o menos* o en otros términos más ambiguos. Estos son los modelos que se utilizan en sistemas de recomendación.

La mayoría de los modelos de confianza ignora completamente la *desconfianza*, o la considera como el final de una misma escala. En [**?**] se propone un modelo que usa una dupla (t, d)

con un grado *t* de confianza y *d* de desconfianza. Para obtener el valor final de confianza se le resta *d* a *t*.

En redes de confianza que existen en la web, lo que ocurre la mayoría de las veces es que para un usuario específico, el resto de los usuarios son desconocidos. Sin embargo hay casos en es útil saber si se puede confiar en un desconocido, y en ese caso, en qué grado es esto posible. Por ejemplo, un usuario desea comprar un producto en un sitio como Amazon, pero se da cuenta de que el producto está mal calificado por otros usuarios. ¿ Son estos comentarios dignos de confianza o no?.

Existen métricas de confianza [**?**] que calculan un estimado de cuánto confiaría un usuario en otro, basado en relaciones de confianza existentes entre otros usuarios en la misma red. Estas métricas usualmente incorporan técnicas basadas en la suposición de que la confianza es, de alguna manera, transitiva. A estas técnicas se les llama *estrategias de propagación*.

En la figura siguiente es posible ver un ejemplo:

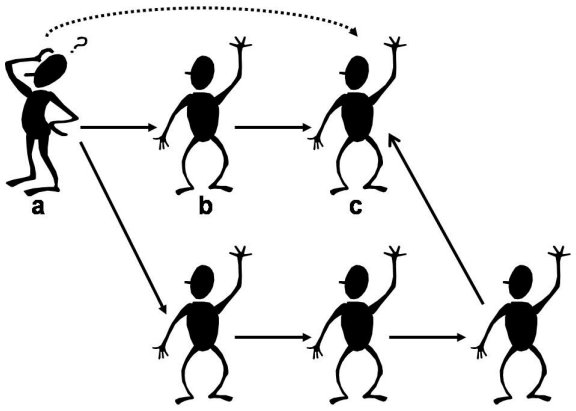


Figura 2.1: Una persona *a* evalúa si confiar o no en una persona *c*

Si una persona *a* desea saber si confiar en *c* y sabe que *b* confía en *c*, se puede decir que *a* tiene al menos cierto grado de confianza en *c*, por transitividad. La estrategia mas básica de propagación se conoce como propagación atómica directa. Este es un modelo que no tiene en cuenta el contexto donde se da la relación de confianza y además usa relaciones transitivas asumiendo que todas las confianzas interpersonales son iguales.

En modelos probabilísticos de confianza, la forma de propagarse es mediante la multiplicación simple, el cual también se use en técnicas graduales, sin embargo éstas poseen un espectro mas amplio de operadores de propagación.

En [**?**] se entrega un ejemplo de propagación de confianza:

Suponer el uso de duplas (t, d) con valores de confianza y desconfianza entre 0 y 1, denotando la gradualidad de los atributos y también permitiendo expresar falta de información como un 0.

- (t_1, d_1) es la confianza que tiene el usuario a sobre el usuario b
- (t_2, d_2) es la confianza que tiene el usuario b sobre el usuario c

Se puede calcular la confianza de a en c como:

$$(t_3, d_3) = (t_1 * t_2, t_1 * d_2) \tag{2.1}$$

Esta estrategia de propagación refleja la actitud de valorar información de fuente confiables y descartar información de personas desconocidas.

Alternativamente, se puede calcular la confianza como:

$$(t_3, d_3) = (t_1 * t_2 + d_1 * d_2 - t_1 * t_2 * d_1 * d_2, t_1 * d_2 + d_1 * t_2 - t_1 * d_2 * d_1 * t_2) \tag{2.2}$$

Esta estrategia hace asume que una fuente desconfiada está dando información errónea maliciosamente. Por lo tanto a confiará en c si es que b es confiable. Si desconfía de b, a hará lo contrario de lo que b sugiere respecto a c. En otras palabras, es como el dicho *El enemigo de mi enemigo es mi amigo.*

Estos ejemplos muestran diferencias entre los autores de la literatura en cuanto al valor que se le da a una fuente confiable y qué hacer ante un agente que no se vea confiable.

Una métrica de confianza también debe incluir estrategias de sumatorias, ya que en una red lo suficientemente grande siempre habrá más de un camino desde *a* hasta *c*.

En estos casos, se pueden utilizar operadores de grafos, tales como camino mínimo, máximo, media aritmética, o promedio ponderado.

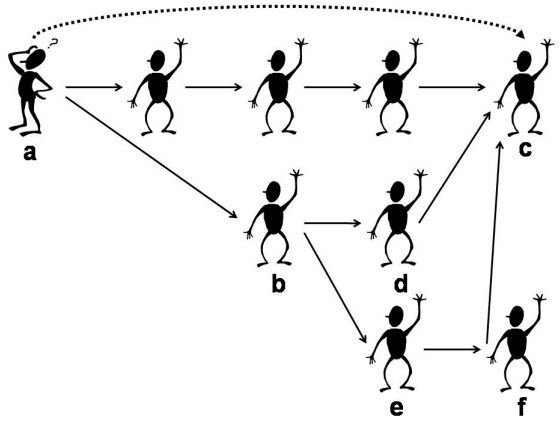


Figura 2.2: Varios caminos para llegar desde *a* hasta *c*

Los operadores ponderados ofrecen la posibilidad de discriminar caminos y seleccionar aquellos que sean de mayor importancia, lo que entrega mayor flexibilidad para el diseño de una red de confianza.

Las estrategias de propagación y sumatorias se pueden combinar de distintas formas, obteniendo distintos resultados, según la importancia que se les de a los caminos y arcos del grafo. Incluso el resultado es distinto si se hace una sumatoria antes de la propagación, para primero elegir un camino entre *a* y *c* y no tener que consultar los valores de confianzas de cada nodo (Que puede llegar a ser altamente costoso con una red lo suficientemente grande), o incluso por razones de seguridad, evitando exponer los grados de confianza,desconfianza de cada nodo.

2.1.2. Sistemas de Recomendación

El sistema de recomendación es el segundo pilar sobre el que se sostiene el sistema que se propone para esta memoria.

	Rapido y Furioso	Los 33	Capitán América	Intensa-Mente	Los Vengadores	50 Sombras de
A			1	5	1	5
B		2	5		4	1
C	1			2	5	1
D	5	2				3
E	2	2	2	2	2	2

Figura 2.3: Las estrellas que dan los usuarios A, B, C y D a distintas películas

Un sistema de recomendación es aquel que intenta predecir una respuesta del usuario hacia opciones.

Los sistemas de recomendación se pueden agrupar en varias categorías, sin embargo las más comunes son [**? ?**]

- Sistemas basados en contenidos. Son aquellos que examinan las propiedades del ítem recomendado. Por ejemplo, si un usuario de Netflix mira muchas películas del género *terror*, entonces el sistema le recomendará una película de terror
- Filtrado Colaborativo. Son sistemas que recomiendan ítems basado en la similitud entre usuarios y/o ítems. Los ítems recomendados a un usuario son los que prefieren usuarios similares

En un sistema recomendador, se pueden identificar dos clases de entidades, *usuarios* e *ítems*[CITA]. Los datos en s[i se representan en una *Matriz de Utilidad*, que da a cada dupla usuario-item un valor que representa el grado de preferencia (Algo así como las *estrellas* en diversos sitios que permiten clasificar películas) con números que van desde el 1 al 5. Se asume que la matriz no estará llena, por lo que la mayoría de las entradas estará en estado desconocido (No hay información sobre la preferencia del ítem).

Teniendo la matriz, el objetivo ahora es predecir los puntajes que habrá en los espacios en blanco. Por ejemplo, el usuario C no ha visto Capitán América, pero tiene gustos similares al usuario A, ya que ambos vieron una película del mismo género (Los Vengadores) y

ambos expresaron preferencia por dicha película, sería razonable asumir que al usuario C probablemente le gustaría ver Capitán América.

Es importante notar que no es necesario llenar todos los vacíos en esta matriz, basta con llenar los que vayan a tener los ratings más altos.

Ya que una de las bases de un sistema de recomendación es encontrar usuarios similares entre sí, hay que mencionar las formas en que se puede medir la similaridad de dos individuos.

Una de las formas consiste en determinar el índice Jaccard entre usuarios, siguiendo con la matriz del ejemplo, tomando las preferencias que tiene cada usuario con las películas como si fueran conjuntos de muestreo.

El índice se calcula como:

$$I_j = \cap / \cup \tag{2.3}$$

Donde $\cap$ es la intersección entre los conjuntos (los elementos que tienen en común) y $\cup$ es la unión entre los conjuntos (la suma de todos los elementos que están en los dos conjuntos). Esta técnica ignora el puntaje que asigna cada usuario a cada película y sólo se preocupa de las películas que ambos hayan visto.

Dicho esto, revisando los ejemplos, comparando los usuarios C y D, se puede ver que su unión es 5 y su intersección es 1. Por lo tanto $1/5 = 0,2$, su índice Jaccard es bajo y efectivamente, es porque sólo han visto una película en común.

Un caso de similaridad es de B y C, que tienen un índice Jaccard de $3/6 = 0,5$.

Una forma de tener en cuenta las preferencias de los usuarios a la hora de determinar si son similares o puede ser redondeando los valores hacia 0 o 1, como intentando aproximar si a la persona le gustó o no la película. Si se reemplazan los ratings de 3, 4 y 5 estrellas por un 1, se dice que al usuario le gustó la película (Siendo optimistas y asumiendo que el valor medio, 3, también dice que a la persona le gustó la película). De lo contrario, se cambia el rating por un 0.

	Rapido y Furioso	Los 33	Capitán América	Intensa-Mente	Los Vengadores	50 Sombras de
A			0	1	0	1
B		0	1		1	0
C	0			0	1	
D	1	0				1
E	0	0	0	0	0	0

Figura 2.4: Se cambiaron los 3,4 y 5 por 1, y los 1 y 2 por 0

Ahora el índice Jaccard para C y D es $0/3 = 0$ y para B y C es $1/2 = 0,5$.

Por último es posible utilizar la correlación de Pearson [**? ?**]. Lo primero es normalizar los valores, restándoles a cada uno su promedio. Así, se convierten bajos ratings en números negativos y altos en positivos.

	Rapido y Furioso	Los 33	Capitán América	Intensa-Mente	Los Vengadores	50 Sombras de
A			-1	3	-1	3
B			3	0	2	-1
C	-0.5	-1.5	-1.5	0.5	3.5	-0.5
D	1	0				1
E	0	0	0	0	0	0

Figura 2.5: La matriz de ejemplo con valores normalizados

Luego se calcula la similaridad como el seno del ángulo entre los vectores que representan a cada usuario. Por ejemplo, para calcular la similaridad entre el usuario B y el C se hace el siguiente cálculo:

$$\frac{3 * -1{,}5 + 2 * 3{,}5 + -1 * -0{,}5}{\sqrt{3^2 + 2^2 + (-1)^2} * \sqrt{(-0{,}5)^2 + (-1{,}5)^2 + (-1,5)^2 + o{,}5^2 + 3,5^2 + (-0{,}5)^2}} = 0{,}3785 \quad (2.4)$$

Esta técnica de filtrado colaborativo toma en cuenta las películas que ambas personas vieron y los ratings que entregan a cada una.

Cuando ya se tiene la correlación entre dos personas, se puede predecir la preferencia de un usuario u por un ítem i con la siguiente fórmula, según [**?**]

$$pred(u,i) = \overline{r} + \frac{\sum\limits_{vecindario} Sim(u,n) * (r_n i - \overline{r_n})}{\sum\limits_{vecindario} Sim(u,n)} \tag{2.5}$$

Donde $Sim(u,n)$ es la similitud entre el usuario u y el usuario n, $\overline{r}$ es el rating promedio (de todos los usuarios) , $r_n i$ es el rating que da el usuario n al ítem i y $\overline{r_n}$ es el rating promedio que entregó el usuario n.

2.1.3. Mejorando un Sistema de Recomendación con una Red de Confianza

Según [CITA], a pesar de recibir mejoras significativas, los sistemas de recomendación siguen teniendo problemas importantes, en especial los de filtrado colaborativo.

Lo que ocurre es que los usuarios generalmente entregan ratings de una fracción minúscula de los ítems disponibles, por lo tanto, los algoritmos tienen problemas para hallar buenos vecindarios y la calidad de las recomendaciones puede sufrir a causa de esto.

También es difícil generar recomendaciones para usuarios nuevos en un sistema (usuarios que partan en frío en una plataforma), ya que no han calificado un número suficiente de ítems y por lo tanto es difícil encontrar usuarios similares.

Ya que los sistemas de recomendación se utilizan en sitios de comercio es frecuente que usuarios mal intencionados abusen del sistema de calificaciones para intentar mejorar el posicionamiento de un producto en el algoritmo de recomendación.

Por último, [**?**] sostiene que las personas tienden a confiar más en recomendaciones hechas por personas de confianza más que por usuarios anónimos similares. Por lo tanto se justifica el uso de redes de confianza para mejorar un sistema de recomendaciones.

En la vida real, una persona que quiera evitar un mal negocio consulta primero a un amigo

y si dicho amigo no sabe, él consulta a un amigo suyo, así sucesivamente, hasta que alguien que aparezca alguien que conozca sobre el tema (un recomendador).

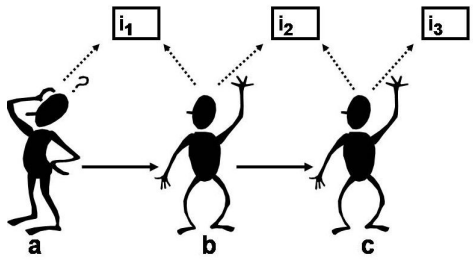


Figura 2.6: Relaciones de confianza en sistemas recomendadores

En la figura 6 se describe una situación donde la líneas sólidas denotan la relación de confianza entre usuarios. En un escenario sin una red de confianza, un filtrador colaborativo no podrías generar una predicción para el ítem i_3 para el usuario a, por falta de información. Pero usando las relaciones de confianza se puede usar una estrategia de propagación que relacione la confianza de a en c y según eso predecir el rating que el usuario a daría al ítem i_3. Por ejemplo, si el usuario a confía en c (por transitividad) y al usuario c le gusta el ítem i_3, entonces se podría decir que probablemente al usuario a también le guste dicho ítem.

Además de resolver el problema de escasez de información, este acercamiento también puede aliviar el problema con usuarios que comiencen en frío en la plataforma.

También, al usar redes de confianza que apoye al sistema de recomendación se puede minimizar el efecto de usuarios malintencionados.

Los sistemas de recomendaciones basados en confianza en dos categorías, según la forma en que se obtienen los valores de confianza [**?**]

- Sistemas que obtienen los valores de confianza directamente desde inputs de usuarios
- Sistemas que infieren valores de confianza, basándose en el historial de un usuario o en reglas de transitividad de usuario a usuario

Los sistemas que obtienen la información de confianza directamente usan estos valores para

determinar cuáles opiniones deberían ser ponderadas más o menos, hay distintas formas de hacer esta ponderación.

Anteriormente, en la ecuación (5) se llega a un rating predicho. Si la confianza entre un usuario a y un usuario b se puede expresar como $t_{u,b}$ (Recordar que este valor puede ir entre 0 y 1). Al multiplicar $t_{u,b}$ por el valor predicho $pred(u,i)$, se puede disminuir el rating esperado, si viene de un usuario poco confiable.

$$pred_conf(u,i) = \frac{\sum_{vecindario} t_{u,b} * pred(u,i)}{\sum_{vecindario} t_{u,b}} \tag{2.6}$$

En este caso, el vecindario son todos los usuarios que opinaron sobre el ítem i.

Volviendo a la ecuación (5), existe un método llamado "Trust-based collaborative filtering"[2] que consiste en reemplazar la similitud entre usuarios $Sim(u,n)$ por la confianza entre éstos $t_{u,n}$, ya que según [CITA] la confianza y la similaridad están correlacionados.

$$pred(u,i) = \overline{r} + \frac{\sum_{vecindario} t_{u,n} * (r_n i - \overline{r_n})}{\sum_{vecindario} t_{u,n}} \tag{2.7}$$

También [**?**] propone limitar el vecindario a usuarios que tengan un nivel de confianza por sobre un mínimo establecido. Esto se conoce como Trust-based filtering.

Existen alternativas también para calcular el valor $t_{u,b}$, una de ellas, conocida como TidalTrust [CITA] consiste en:

- Permitir sólo caminos cortos entre el usuario a y el c, ya que son los que entregan resultados más precisos [CITA], para ésto se determina un largo variable para el camino y se busca el más corto
- Permitir sólo información que provenga de usuarios con un nivel mínimo de confianza. Éste se determina cómo el valor máximo del peso de un camino (El peso de un camino está determinado por el valor más bajo de confianza)

[2]Filtrado Colaborativo, basado en confianza.

- Por lo tanto, el vecindario en la ecuación (5) queda definido como todos aquellos usuarios que superen el mínimo establecido anteriormente

Otra técnica se llama MoleTrust. Es una idea similar a TidalTrust, se diferencia en que obtiene valores de confianza desde todos los caminos que lleven desde el usuario a al usuario c, con un largo límite definido d. La otra diferencia con TidalTrust es que también se define manualmente el nivel de confianza mínimo a aceptar (En [CITA] se definen valores de confianza aceptables desde 0,6 hasta 1).

El problema con los algoritmos descritos anteriormente es que requieren que los usuarios describan explícitamente el nivel de confianza con otros usuarios. Ésto puede no ser siempre práctico, por lo que existen algoritmos que pueden inferir estimaciones de valores de confianza, sin información explícita, según el comportamiento histórico de cada usuario, una persona que haya dado opiniones valoradas por la comunidad será más confiable para el resto de los usuarios, en general.

Una forma es implementar un sistema de reputación, que [**?**] describe como confianza a nivel de perfil y de ítem.

Para esto, se genera una predicción $p_{a,i}$ del rating que el usuario u colocaría sobre el ítem i en el que está interesado el usuario objetivo a, si la diferencia está dentro de un rango ϵ con respecto al rating que el usuario realmente puso, se dice que u acertó.

Así, se calcula la confianza $t^P_{a,u}$ a nivel de perfil desde a hacia u como el porcentaje de veces que el usuario u acertó.

También se puede calcular la confianza a nivel de ítem $t^P_{a,u}$ como el porcentaje de predicciones sobre el ítem i acertadas.

[**?**] Concluye al comparar los sistemas de recomendación con y sin redes de confianza entregan resultados similares, sin embargo, los sistemas sin redes de confianza tienen peor rendimiento en ítems que tienen opiniones muy contradictorias.

Capítulo 3

Propuesta

Se propone el desarrollo de un sistema que incorpore elementos, tanto de la computación basada en ”trust” como las de web social, más específicamente, los sistemas recomendadores. El dominio escogido para este sistema es el turismo en la región metropolitana. Actualmente existe una gran variedad de centros culturales y otros lugares interesantes para visitar, sin embargo, un visitante no tiene tiempo para visitarlos todos y además puede tener preferencia sobre algún tipo de lugar por sobre otros. La idea es, primero, ofrecer una plataforma en la que un usuario pueda encontrar lugares culturales cercanos y poder dejar una reseña escrita que describa brevemente la experiencia de la visita. Por otra parte, se espera que las opiniones de la comunidad que use este sistema permitan predecir los lugares que serán del gusto del usuario objetivo.

3.1. Características

Las principales características que incluye el sistema ”VISIT CHILE” son 8, entre todas abarcan conceptos de web social, geo-localización y ”trust computing”.

1. Posibilidad de dejar opiniones en forma de reviews”que consisten en una calificación, del 1 al 5 del lugar visitado, junto con una descripción corta del lugar

2. Visualización de perfil de usuario, con avatar y lista de reviews.[es]critas
3. Visualización de lugares culturales, que incluyen una fotografía del lugar, una descripción y además, la lista de opiniones de los usuarios al respecto
4. Visualización de lista de lugares recomendados para un usuario, ordenados según el puntaje predicho
5. Funcionalidad que permite calificar las reseñas que deja cada usuario, indicando si le fueron útiles o no
6. Funcionalidad de tags para cada lugar cultural, según las actividades que se pueden realizar ahí
7. Funcionalidad de buscar lugares según el tag que posean
8. Funcionalidad de buscar lugares según la cercanía geográfica

3.2. Arquitectura de la solución

El sistema ”CHILEEE” consiste en una aplicación web desarrollada utilizando el framework Ruby on Rails, el cual utiliza el paradigma MVC (Modelo, Vista, Controlador). Para almacenar todos los datos (usuarios, lugares, reseñas y trust) Se utilizó Neo4j, una base de datos NoSQL diseñada para almacenar grafos. Neo4j permite unificar la lógica de trust y la del sistema de recomendación, diseñando el sistema completo como un grafo. En este grafo existen tres tipos de nodos.

1. Los de color azul representan a los usuarios, quienes son los que dejarán sus opiniones respecto a los lugares que visitan.
2. Los de color verde son los ”Items”, en específico, los lugares culturales sobre los cuales se opina
3. Finalmente, los de color rojo son los tags, que describen los lugares, según los tipos de actividades que se pueden hacer en cada uno de ellos.

Todo grafo debe tener, además de nodos, arcos, en este caso, representan las relaciones entre las entidades que describe el sistema. Existen cuatro tipos, trust, vote, tagged y review.

1. La relación review va desde un usuario hacia un ítem, indica la opinion que tuvo el usuario en particular acerca del item al que apunta. Además de la dirección hacia la que apunta, contiene la información sobre la opinion dada, una corta reseña y una puntuacion en estrellas

2. La relación vote describe la opinión de un usario sobre la reseña que haya escrito otro. Ya que una persona no siempre puede estar de acuerdo con la apreciación de otra, se refleja esto en el contenido del arco, que indica si la reseña le fue útil al usuario actual o no.

3. La relación tagged es entre un item y varios tags, cuando un tag apunta a un ítem, significa que éste item ha sido ”tagueado” o ”marcado” por este tag y es por lo tanto descrito, en parte, por éste.

4. La relación trust es entre un usuario y otro, indica el nivel de confianza que uno hacia el otro, el contenido de este arco es un valor que intenta representar el nivel de confianza, un nivel cero significa que no existe confianza, mientras que un nivel 1 indica confianza total. Por defecto el valor asignado es 0.5

3.3. Implementación de las características

En la sección X se hizo un listado de las características que posee el sistema. A continuación se detallará el diseño e implementación de las características anteriormente nombradas.

3.3.1. Estimación de trust

Una de las características que hace posible todo es la de escribir reviews. Un usuario registrado puede acceder a cualquier ítem de la base de datos y, si no lo ha hecho ya, escribir una review sobre el lugar que describe el ítem, junto con su puntaje en estrellas, del 1 al 5.

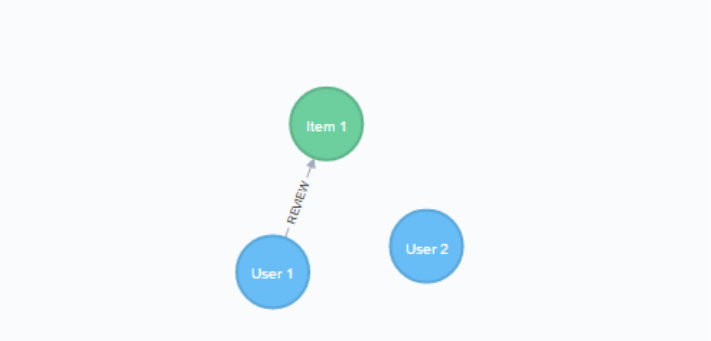


Figura 3.1: Un ejemplo sencillo de los grafos que genera el sistema. Los nodos azules representan usuarios, mientras que los nodos verdes representan los ítems (lugares culturales) a evaluar. El arco que une al Usuario 1 con el Item 1 es la reseña (Review) que le da el usuario al ítem.

Ya que el usario pertenecerá a una amplia comunidad, repartida entre varias comunas de la RM, no es posible garantizar que conozca personalmente a cada otro usuario del sistema. Además, dado que la información disponible sobre otras personas es limitada, ya que la única forma de interacción entre usuario que se permite es la escritura de reviews, se hace apropiado que se estime la confianza según el nivel de acuerdo que exista entre las opiniones de los usarios. Por lo tanto, se implementó un sistema que permite valorar como útil o poco útil la opinión de otro usuario. El usuario puede hacer ésto al hacer clic en la mano con dedo pulgar hacia arriba o la que apunta hacia abajo.

Para ejemplificar, se utilizarán 3 usuarios de ejemplo: Usuarios A y B

Suponer que el Usuario A ingresa a la vista de un item arbitrario, dentro del cual hay sólo una reseña, escrita por el usuario B. Ahora imaginar que el usuario A está de acuerdo con lo que el usuario B escribió y decide valorar positivamente su reseña. A partir de ésto hay dos casos posibles:

1. Si el usuario A y el usuario B no tenían relación de trust existente, se crea una entre ambos, con el valor por defecto, 0.5
2. Si el usuario A y el usuario B ya tenían una relación de trust, ésta se refuerza y se aumenta el valor de trust de A hacia B en 0.1. Si la valoración hubiese sido negativa,

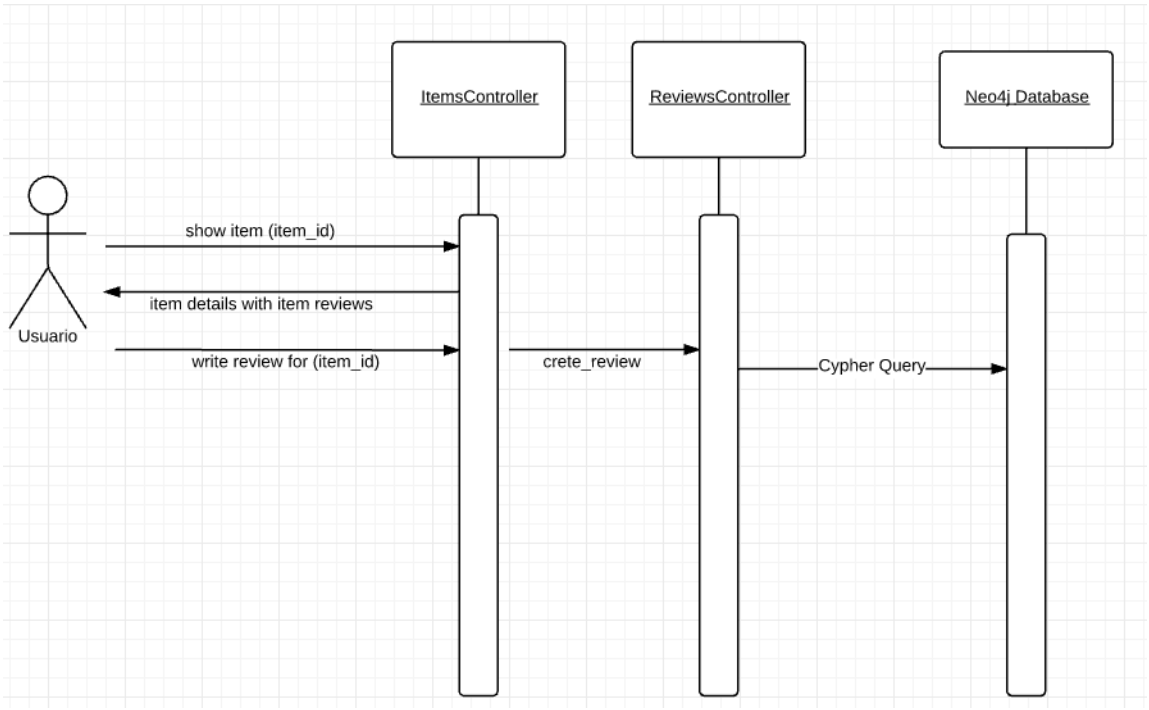


Figura 3.2: Diagrama de interacción entre el usuario, los controladores y la base de datos

se habría restado 0.1 al trust de A a B.

De esta manera, se pretende un acercamiento a como funciona la confianza en la vida real, no de forma binaria (confiar o no confiar), sino más bien, de forma gradual.

3.3.2. Estimación de ítem recomendado

Uno de los pilares fundamentales del sistema es el algoritmo que permite obtener los ítems recomendados. Para ésto, después de evaluar las opciones posibles (ver estado del arte) se decidió por una variante del algoritmo collaborative filtering (encontrar nombre exacto) que incluye el factor de confianza como variable a considerar. En palabras simples, el sistema intenta predecir el puntaje que el usuario objetivo dará a cierto ítem en partícular, los ítems que reciban mayor puntaje serán los recomendados. A continuación se explica en detalle su funcionamiento.

Cuando un usuario ingresa al sistema usando sus credenciales, el controlador de usuarios

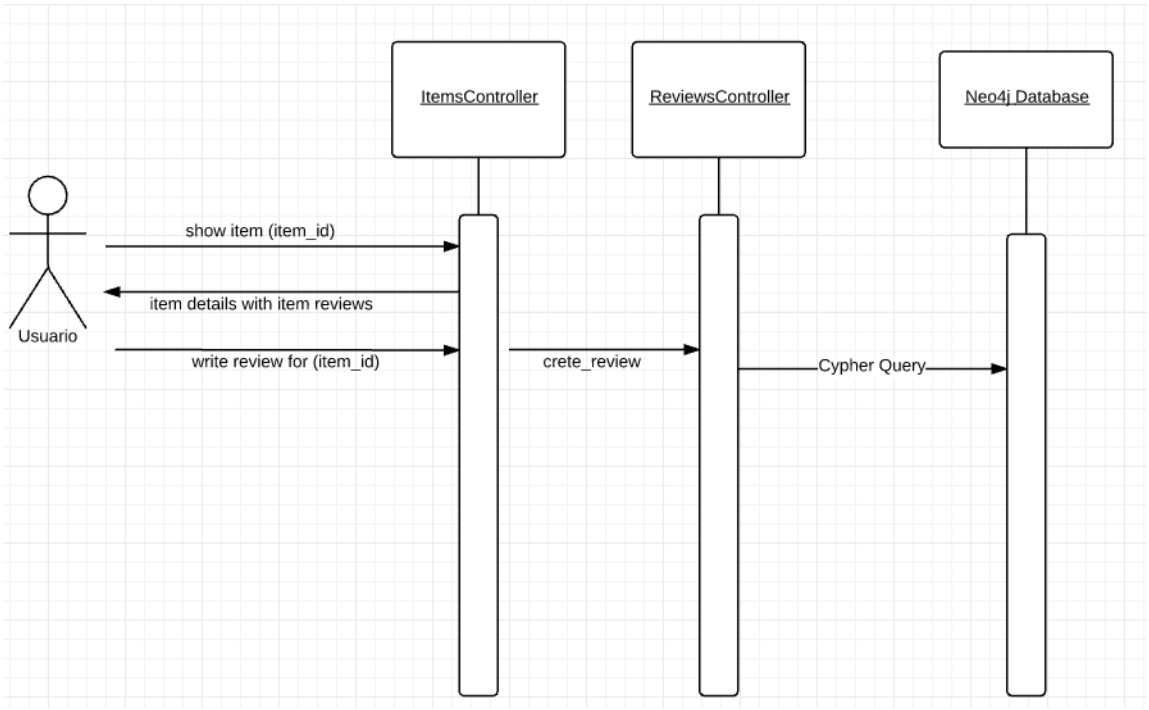


Figura 3.3: Diagrama de interacción entre el usuario, los controladores y la base de datos

hace una redirección hacia la función de recomendar. La cual calcula los puntajes predichos para todos los ítems candidatos con la siguiente fórmula

$$pred(u,i) = \overline{r} + \frac{\sum\limits_{vecindario} Sim(u,n) * (r_n i - \overline{r_n})}{\sum\limits_{vecindario} Sim(u,n)} \tag{3.1}$$

Donde $Sim(u,n)$ es la similitud entre el usuario u y el usuario n, $\overline{r}$ es el rating promedio (de todos los usuarios) , $r_n i$ es el rating que da el usuario n al ítem i y $\overline{r_n}$ es el rating promedio que entregó el usuario n.

La función ”recommend”

Esta es una función que no recibe parámetros, funciona de la siguiente forma.

Cuando el usuario se autentica en el sistema, el controlador de usuarios llama a la función

”recommend”, la cual empieza el proceso de recomendacion. Cada proceso de recomendación tiene un ”consumidor” y varios ”productores”.

El usuario que recibe la predicción es el que se conoce como ”consumidor”, mientras que los usuarios que que entregan la información para cada predicción son los ”productores”

Lo primero que hace este algoritmo es consultar por todos los ítems que se tengan registrados en la base de datos, si este arreglo no está vacío, se llama a todas las reviews que tenga el consumidor y además se guarda un arreglo con los ítems que tengan reviews del consumidor.

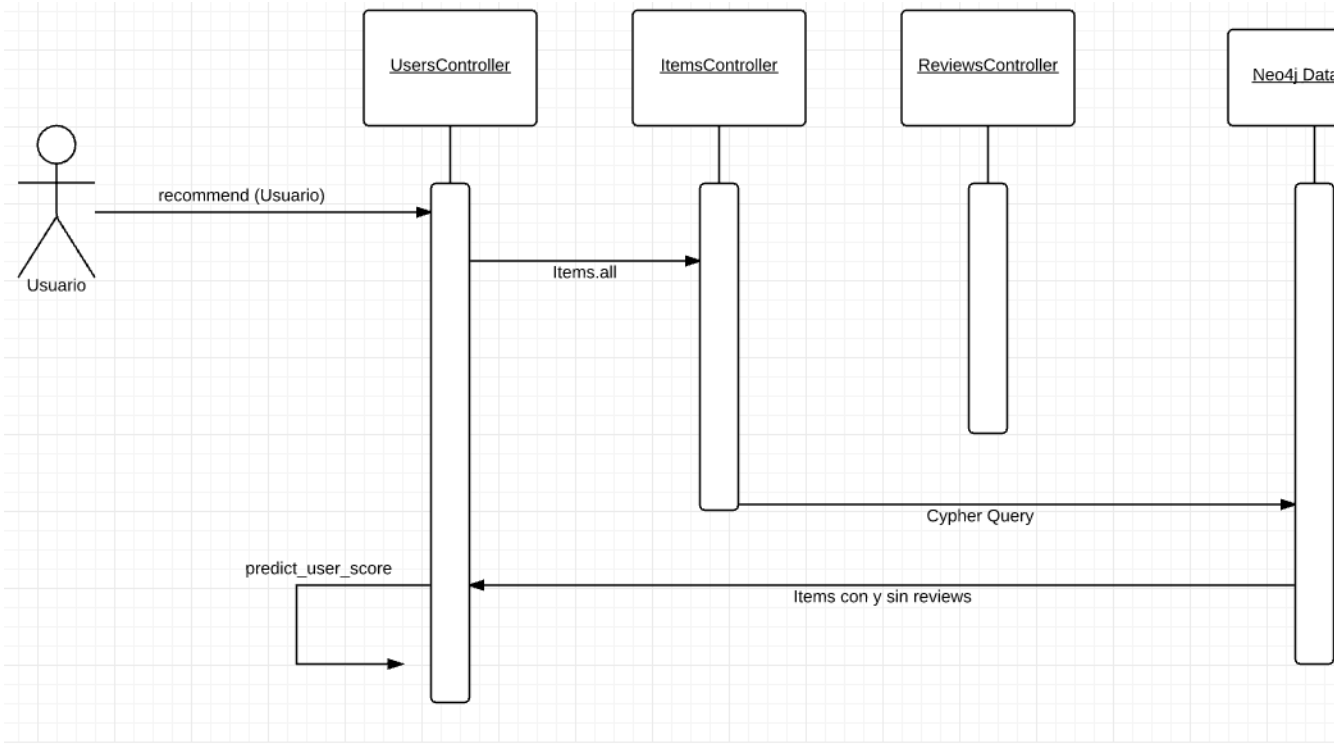


Figura 3.4: Diagrama de interacción entre el usuario, los controladores y la base de datos

A continuación, se hace un arreglo con todos los lugares que el consumidor no ha visitado, ” not reviewed ”

Por cada elemento que haya en el arreglo ” not reviewed ” se calculará el puntaje predicho para el consumidor, usando la función ” predict_user_score ”.

$$producers = [[user1, item1, stars1], [user1, item2, stars2], ...] \quad (3.2)$$

Figura 3.5: Diagrama de interacción entre el usuario, los controladores y la base de dato

La función "predict_user_score"

Esta es una función que recibe como parámetro el ítem al cual se quiere predecir el puntaje a ser dado por el usuario actual (consumidor).

El sistema debe, entonces encontrar todas las reviews que hayan sido escritas sobre el ítem objetivo y determinar además quienes serán los productores para esta sesión.

Para comparar las reviews del consumidor con cada productor, se crea un arreglo de productores con la siguiente estructura:

Capítulo 4

Implementación

Para facilitar la comprensión del funcionamiento del algoritmo, se va a usar el siguiente ejemplo:

En este ejemplo, el usuario que entra al sistema y cuyos puntajes son predichos se conoce como consumidor, en este caso es el Usuario 1.

Cada ítem para el cual el consumidor aún no ha escrito una review y que se encuentre a cierta distancia geográfica es un potencial ítem recomendado, pero antes es necesario conocer qué puntaje le asignaría el consumidor.

El proceso de calcular el puntaje predicho para cada ítem se conoce como sesión. Como se puede notar, cada ítem candidato implica una sesión de recomendación.

Comenzando por el primer ítem candidato, al iniciar la sesión, el sistema consulta por todas las reviews que hayan sido escritas sobre el ítem actual. Los usuarios que escribieron estas

	Plaza de Armas	Cerro Sta. Lucia	Cerro San Cristobal	Museo de Bellas Artes	Quinta Normal	Teatro Municipal	Centro GAM
Juan	5	4	2		1		
Ignacio	3		2		5		4
Perla	3	3		1	4	3	
Inés	5	5		1	2		5
Sam	1					4	3

Figura 4.1: Las estrellas que dan los usuarios A, B, C y D a distintas películas

	Plaza de Armas	Cerro Sta. Lucia	Cerro San Cristobal	Museo de Bellas Artes	Quinta Normal	Teatro Municipal	Centro GAM
Juan	5	4	2		1		
Ignacio	3		2		5		4
Perla	3	3		1	4	3	
Inés	5	5		1	2		5
Sam	1					4	3

Figura 4.2: La misma tabla de la figura x, imaginar que se quiere conocer el puntaje que Juan dará al Museo de Bellas Artes, para eso se deben tener en cuenta sólo los usuarios que hayan visitado este lugar, o sea, Perla y Inés.

	Plaza de Armas	Cerro Sta. Lucia	Cerro San Cristobal	Museo de Bellas Artes	Quinta Normal	Teatro Municipal	Centro GAM
Juan	5	4	2		1		
Ignacio	3		2		5		4
Perla	3	3		1	4	3	
Inés	5	5		1	2		5
Sam	1					4	3

Figura 4.3: Tabla x, pero destacando que para el cálculo del puntaje al Museo de Bellas artes, sólo se tomaran en cuenta lugares que hayan sido visitados por los tres usuarios, Juan, Perla e Inés.

reviews se conocen como productores.

De este grupo de productores, se escogen sólo los que hayan escrito reseñas sobre al menos dos ítems sobre los que el consumidor haya también dado su opinión.

En la sección de estado del arte se describieron varias formas en que se pueden integrar los algoritmos de recomendación con los de trust. Para este dominio en particular, ya que se intenta hacer un acercamiento al comportamiento observado en la vida diaria entre personas, se decidió que se utilizará el trust como un umbral, en el que la opinión de una persona ”no confiable” se considera inválida y por lo tanto, es ignorada.

Por lo tanto, después de determinar los posibles usuarios productores, existe otro filtro que determina quiénes serán de confianza. Después de obtener la lista inicial de productores, se determina el nivel de trust que tiene el consumidor con cada uno de los productores. Si no existe una relación directa de trust entre el consumidor y el posible productor, ésta se calcula mediante ”tidal trust”. Una vez que se tiene el valor de trust, el productor candidato es removido de la lista si es que su trust con el consumidor es menor a 0.5, en caso contrario,

$$sim(u,n) = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(n_i - \bar{n})}{\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2} + \sqrt{\sum_{i=1}^{n}(n_i - \bar{n})^2}} \tag{4.1}$$

Figura 4.4: La similitud entre dos conjuntos, según la correlación de Pearsson, en este caso, interpretar los valores de u como las estrellas que ha dado Juan a los lugares que visitó y los de n los que ha dado Perla.

permanece en la lista.

Luego, para estimar numéricamente el parecido de las opiniones del consumidor con cada productor, se utiliza la correlación de Pearsson (cita). La idea detrás de esta lógica es que mientras más parecido sea el consumidor con un productor, más peso tendrá éste sobre la estimación del puntaje predicho. Empezando por Juan y Perla:

Conociendo el puntaje que ha puesto en promedio cada productor, $\bar{u}$ para Juan y $\bar{n}$ para Perla, se calcula la diferencia con el puntaje dado. Reemplazando con los promedios y los valores de reseñas, se tiene:

$$\frac{3 * -1{,}5 + 2 * 3{,}5 + -1 * -0{,}5}{\sqrt{3^2 + 2^2 + (-1)^2} * \sqrt{(-0{,}5)^2 + (-1{,}5)^2 + (-1,5)^2 + o{,}5^2 + 3,5^2 + (-0{,}5)^2}} = 0{,}3785 \tag{4.2}$$

Este proceso se repite para cada dupla de consumidor-productor posible.

Con esto, se tiene el numerador de la ecuación (referencia), el denominador es el producto de las sumatorias de las desviaciones estándares, para Juan y para Perla.

Con esto, se llega finalmente a un puntaje estimado, que dará el consumidor al ítem actual. Este proceso se repite para cada ítem candidato. Luego, cuando ya se tienen todos los puntajes predichos, se considera como recomendables sólo los ítems cuyo puntaje estimado sea al menos 3 estrellas.

Son estos los ítems que el sistema va a mostrar en la vista de recomendados, ordenados de mayor a menor puntaje.

	Museo de Bellas Artes	Teatro Municipal	Centro GAM
Puntaje Predicho Juan	1.28	3.60	3.09

Figura 4.5: Se pueden ver los puntajes que el sistema predice para Juan, se puede ver que el sistema estima que a Juan no le va a gustar el Museo de Bellas Artes, sin embargo, se puede ver que dará puntajes por sobre 3 estrellas al Teatro Municipal y al Centro GAM, por lo tanto se les recomienda esos lugares a Juan.

Conclusiones