Ingeniería Investigación y Tecnología
volumen XX (número 3), julio-septiembre 2019 1-12
ISSN 2594-0732 FI-UNAM artículo arbitrado
Información del artículo: Recibido: 15 de agosto de 2017, reevaluado: 8 de enero de 2018 y 14 de marzo de 2019, aceptado: 13 de mayo de 2019

http://dx.doi.org/10.22201/fi.25940732e.2019.20n3.025

Evaluación de un sistema de recomendación híbrido de trabajos de titulación
Evaluating a hybrid recommender system of theses

Vences-Nava Rodrigo
Universidad Autónoma de Yucatán
Facultad de Ciencias Antropológicas
Correo: vnava@correo.uady.mx
https://orcid.org/0000-0001-8577-8900

Menéndez-Domínguez Víctor Hugo
Universidad Autónoma de Yucatán
Facultad de Matemáticas
Correo: mdoming@correo.uady.mx
https://orcid.org/0000-0003-3587-1263

Medina-Peralta Salvador
Universidad Autónoma de Yucatán
Facultad de Matemáticas
Correo: mperalta@correo.uady.mx
https://orcid.org/0000-0003-4472-6690

Resumen

La tarea de evaluar un sistema de recomendación no se limita a verificar que el algoritmo de recomendación funcione correctamente, sino que debe contemplar además la validación de que cumple con los objetivos para los cuales fue diseñado. Esto último varía dependiendo del sistema de recomendación, ya que para los sistemas comerciales significaría aumentar sus ventas, pero para uno del sector educativo podría ser mejorar el aprovechamiento de los estudiantes. Este artículo se enfoca en la evaluación de un sistema de recomendación, que presenta al usuario, trabajos de titulación similares a los que ha consultado anteriormente con base en diferentes fuentes de información asociadas a los mismos (metadatos) y su interacción con el sistema y la comunidad de usuarios. Se realizaron tres tipos de estudios para evaluar la precisión del motor de búsqueda y recuperación, el uso del filtrado de palabras clave en la recomendación colaborativa y por último la utilidad, facilidad y percepción del sistema por medio de los cuestionarios *Escala de Usabilidad del Sistema* y el *Modelo de Aceptación Tecnológica*. Los análisis estadísticos comprueban la efectividad y aceptación del Sistema de Recomendación por parte de los tres grupos de usuarios participantes en los experimentos (profesores, tesistas y alumnos en general).
Descriptores: Sistemas de recomendación, SUS, TAM, evaluación de sistemas de recomendación.

Abstract

The task of evaluating a recommendation system is not limited to checking that the recommendation algorithm works correctly, but must also consider the validation that it fulfills the objectives for which it was designed. The latter varies depending on the recommendation system, since commercial systems increase sales, but for one of the education sector could be improving student achievement. This article focuses on the evaluation of a recommendation system, which presents to the user, degree jobs similar to those previously consulted based on different sources of information associated with them (metadata) and their interaction with the system and the community of users. Three types of studies were performed to evaluate the search engine's accuracy and retrieval, the use of keyword filtering in the recommendation and the latest utility, ease and perception of the system through the questionnaires System Usability Scale and Technology Acceptance Model. Statistical analyzes verify the effectiveness and acceptance of the Recommendation System by the three user groups in the experiments (teachers, theses's students and students in general).
Keywords: Recommender systems, SUS, TAM, evaluating recommender systems.

Introducción

Un Sistema de Recomendación (SR) recopila información sobre las preferencias de los usuarios sobre un conjunto de elementos (películas, canciones, libros, etcétera), desde diferentes fuentes de información para proporcionar a los usuarios predicciones y recomendaciones de elementos similares, tratando de equilibrar factores como exactitud, novedad, dispersión y estabilidad en las recomendaciones (Bobadilla *et al*., 2013 y Ricci *et al*., 2011).

Para obtener estas recomendaciones personalizadas, el contexto del usuario y el perfil del usuario (preferencias del usuario, histórico de transacciones, datos demográficos) necesitan ser considerados (Bobadilla *et al*., 2013 y Erdt *et al*., 2015).

El proceso de evaluación del software considera realizar dos tipos de valoraciones para tratar de garantizar la calidad de un producto: verificación y validación (Pressman, 2010). Mientras que la verificación se orienta a comprobar que el producto ha sido desarrollado correctamente, la validación evalúa el producto desarrollado contrastándolo con las necesidades planteadas, es decir, si se ha construido el producto correcto según los objetivos iniciales.

Los tipos de metodología de evaluación aplicados para la evaluación de SR son clasificados en cuatro categorías: experimentos offline, estudios de usuarios, pruebas en la vida real y la no evaluación (Erdt *et al*., 2015).

En este trabajo se presentan los resultados de la evaluación de un sistema de recomendación de trabajos de titulación donde se realizan tres estudios y participan tres tipos de usuarios con el objetivo de evaluar la precisión del motor de búsqueda y recuperación, el uso del filtrado de palabras clave y la utilidad, facilidad y percepción del sistema por parte del usuario. El SR evaluado fue el Sistema de Recomendación de Trabajos de Titulación (SIRETT) (Vences *et al*., 2015), el cual recomienda a los usuarios trabajos similares utilizando como métrica de similitud entre documentos, la frecuencia de los términos contenidos en los mismos y considerando además su interacción con la comunidad de usuarios.

La Universidad Autónoma de Yucatán (UADY) cuenta con un Sistema Bibliotecario Institucional (SISBI), sin embargo, este no almacena información histórica alguna respecto a los usuarios y sus consultas, y no hace uso de tal información para generar recomendaciones a los usuarios, por lo cual se espera que SIRETT pueda ser un SR opcional o complementario al SISBI.

La organización del trabajo se presenta a continuación, con los principales criterios para la evaluación del sistema de recomendación. La siguiente sección muestra los resultados obtenidos mediante la aplicación de los estudios. Finalmente, se presentan las conclusiones.

Criterios de evaluación

Wu *et al*. (2012) dividen las métricas de evaluación de los SR en dos clases: criterios de evaluación basados en el algoritmo de recomendación (precisión, cobertura, diversidad, novedad y serendipia), y criterios de evaluación independientes del sistema de recomendación, desde el punto de vista del sistema (confianza, robustez, adaptabilidad y escalabilidad) y desde el punto de vista del usuario (confianza, riesgo, utilidad, privacidad y preferencia del usuario).

Un Sistema de Recuperación de Información (SRI) está estrechamente relacionado con los sistemas de recomendación, ya que estos últimos se desarrollan a partir de la necesidad de personalizar la información recuperada por los motores de búsqueda con el fin de adaptarse a las necesidades del usuario, lo cual implica "interpretar" el contenido de los documentos, tomar en cuenta el perfil del usuario y clasificar los resultados de acuerdo con su relevancia.

La relevancia es una medida de la capacidad del documento recuperado para satisfacer la necesidad de información del usuario con base en la consulta realizada. Por otro lado, la efectividad es puramente una medida de la capacidad del sistema para satisfacer al usuario en términos de la relevancia de los documentos recuperados (Van, 1979).

Una primera distinción hecha en la evaluación de los motores de búsqueda es entre efectividad y eficiencia. La efectividad se define como la medición de qué se define tan bien el ranking producido por el motor de búsqueda corresponde al ranking basado en la relevancia para el usuario. La eficiencia se define en términos de requerimientos de tiempo y espacio para que el algoritmo produzca el ranking (Croft *et al*., 2015).

Otra característica de los SRI es la exhaustividad, intuitivamente expresa de qué tan bien está elaborado el motor de búsqueda para encontrar todos los documentos relevantes para una consulta, y la precisión indica qué tan bien está hecho al rechazar documentos no relevantes. La definición de esta medida asume que hay un conjunto de documentos recuperados y otros que no son recuperados (el resto de los documentos).

En la Tabla 1, A es el conjunto de los documentos relevantes para una consulta, ¬A es el conjunto no relevante (complemento), B es el conjunto de los documentos recuperados y de la misma forma tiene su complemento. $A \cap B$ es el conjunto de los documentos relevantes y recuperados.

Tabla 1. Conjunto de documentos definido por una búsqueda simple con relevancia binaria

	Relevante	No relevante
Recuperado	A ∩ B	¬A ∩ B
No recuperado	A ∩ ¬B	¬A ∩ ¬B

Exhaustividad (1) es la proporción de documentos relevantes que son recuperados y la precisión (2) es la proporción de los documentos recuperados que son relevantes.

$$Exhaustividad = \frac{A \cap B}{A} \quad (1)$$

$$Precisión = \frac{A \cap B}{B} \quad (2)$$

El método más popular para resumir la efectividad de un ranking es promediando los valores del cálculo de precisión de la posición del ranking cuando un documento relevante es recuperado.

Para proporcionar evaluaciones realistas de la efectividad de un algoritmo de recuperación, este debe ser probado en un número de consultas. La medida de efectividad MAP (*Mean Average Precision*) resume la efectividad de ranking de múltiples consultas.

El supuesto básico en un sistema de recomendación es que un sistema que proporcione predicciones más exactas será preferido por los usuarios. La *exactitud* de la predicción es normalmente independiente de la interfaz del usuario y puede ser medida en un estudio no presencial (offline). Si se quiere conocer la predicción de la valoración que un usuario dará a un elemento, la métrica más utilizada es la desviación de la raíz cuadrada media (Ricci *et al.*, 2011).

Cuando se quiere mejorar un sistema es importante conocer la *preferencia del usuario* a un sistema sobre otro. Normalmente es fácil entender cuando se comparan propiedades específicas. Por lo tanto, mientras la satisfacción del usuario es una medida importante, dividir la satisfacción en componentes más pequeños ayuda a entender el sistema y mejorarlo.

En ocasiones puede ser bueno para un sistema recomendar algunos ítems que el usuario previamente conozca y le gusten, ya que así el usuario observa que el sistema proporciona recomendaciones razonables, lo cual incrementa su *confianza* para recomendaciones de ítems desconocidos. Otra forma de incrementar la confianza es explicar las recomendaciones que el sistema provee. Alejandres *et al*. (2016) determinaron que las explicaciones textuales ayudan a incrementar la confianza del usuario hacia el sistema. Una forma de evaluar es preguntando al usuario si las recomendaciones son razonables en un estudio de usuario. También es posible asociar la medición si registramos la cantidad de recomendaciones seguidas o por medio de la repetición de usuarios quienes confían en el sistema y regresan en el futuro.

La *privacidad* en un entorno colaborativo es clave, ya que es importante para muchos usuarios que sus preferencias permanezcan privadas. La privacidad puede ir en detrimento de la precisión de las recomendaciones.

Dentro de las métricas de eficiencia que se definen en términos de requerimientos de tiempo y espacio para que el algoritmo produzca el ranking, se tienen el *rendimiento* (consultas procesadas por segundo), el *tiempo transcurrido de indexación*, el *tiempo de procesamiento del índice* (no cuenta el tiempo de espera de entrada y salida (I/O), la *latencia de la consulta* (medida en milisegundos), el *espacio de indexación temporal* y el *tamaño del índice* (espacio para almacenamiento).

Análisis de resultados

Como se mencionó en la introducción, el sistema que se evaluó fue SIRETT, el cual emplea un modelo de recomendación híbrido basado en fuentes de información asociadas a los trabajos de titulación (metadatos) y su interacción con los usuarios para mejorar la búsqueda, selección y recuperación de trabajos de titulación similares y acordes a las necesidades del usuario; este sistema ofrece una solución relevante al acceso a la información de trabajos de titulación de la Facultad de Ciencias Antropológicas (FCA) de la UADY. La FCA-UADY cuenta con una matrícula de más de 700 alumnos y cerca de 100 profesores quienes son beneficiados por el uso del sistema.

La Figura 1 muestra la arquitectura del sistema y el diagrama de clases. Es una arquitectura cliente-servidor divida en tres niveles: interfaz, servicios (incluye nivel de pre procesamiento) y datos.

El sistema mantiene el repositorio de los trabajos de titulación (411 documentos), las relaciones de similitud entre ellos (cada uno con los otros 410 restantes), los índices invertidos generados a partir de su contenido y además almacena toda interacción del usuario con el sistema. Toda esta información se almacena en ocho tablas relacionadas (Figura 2).

Se realizaron tres pruebas diferentes para validar y verificar el correcto funcionamiento del sistema. La primera prueba se realizó con profesores para evaluar la precisión en la recuperación de información. La segun-

da prueba se realizó con alumnos tesistas, para evaluar el proceso de recomendación por uso. Finalmente, la tercera prueba se realizó con un grupo de alumnos, del cual no forman parte los tesistas, para evaluar la facilidad de uso del sistema en general. Los tres tipos de participantes en las pruebas evaluaron además la utilidad, facilidad y percepción del sistema.

Pruebas del sistema de recomendación híbrido con profesores

Para verificar el correcto funcionamiento del sistema y validar si cumple con los objetivos para los cuales fue diseñado, se dividió en dos partes la evaluación, para lo cual se invitaron a 19 profesores de la FCA-UADY, los cuales fueron seleccionados con base en su conocimiento en buscadores bibliotecarios y de artículos científicos, además de que previamente participaron con sus grupos de alumnos en la fase de liberación del sistema. La mayoría de los integrantes cuenta con más de 10 años de experiencia docente. Las áreas de formación académica a las que pertenecen son Ciencias Sociales y Humanidades. Un 15% solo cuenta con estudios de licenciatura, el resto ha cursado alguna maestría o doctorado. La mayoría (84%) de los sujetos cuenta con al menos 5 años de experiencia en buscadores bibliotecarios y de artículos científicos. Todos cuentan con conocimiento en modelos educativos de formación integral y se encuentran en capacitación constante en temas de pedagogía que son relevantes para su institución.

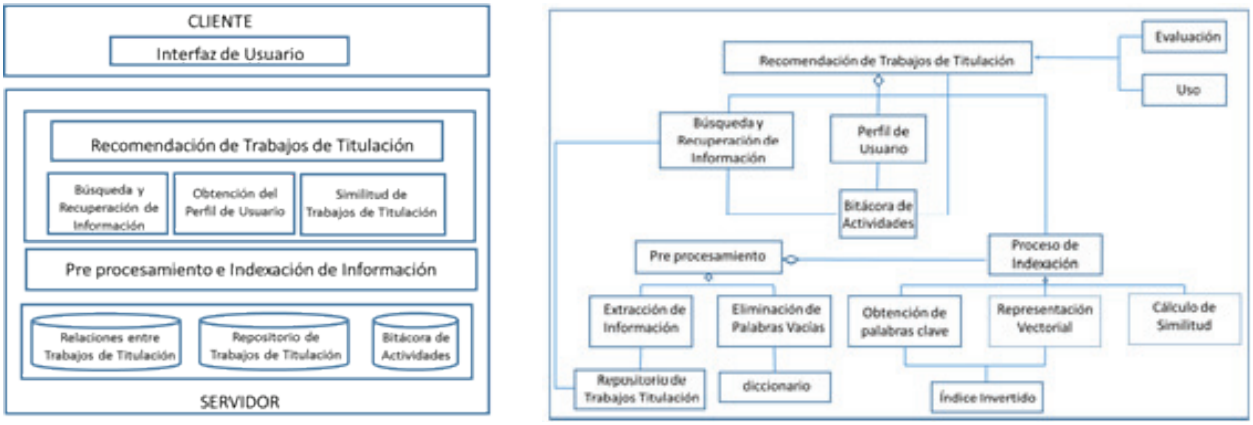

Figura 1. Representación de la arquitectura de SIRETT y el diagrama de clases

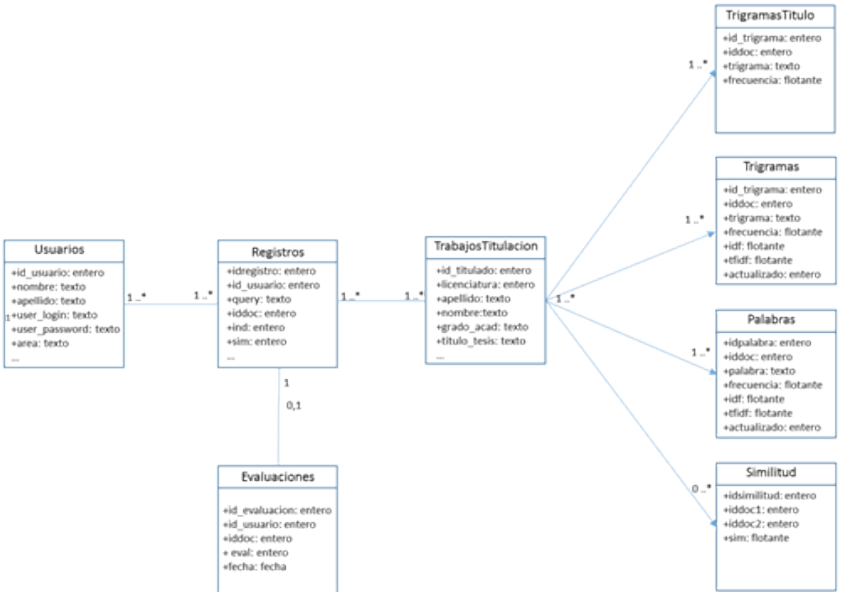

Figura 2. Esquema de las tablas de la base de datos y sus relaciones

Precisión en la búsqueda y recuperación de información

Para la primera parte de la evaluación se realizó una prueba de precisión y para la segunda se analizó la facilidad, utilidad y percepción del motor de búsqueda y recuperación de información del sistema SIRETT.

Existe una métrica en los SRI que se denomina exactitud (accuracy), que en la tabla de contingencia (Tabla 1, sección 2) corresponde a la fracción de esta clasificación que es correcta. Esto parece aceptable, ya que un SRI puede ser pensado como un clasificador de dos clases (relevantes y no relevantes). Hay una buena razón por la que esta métrica no es adecuada para los SRI. En casi todas las circunstancias los datos son extremadamente sesgados: normalmente más de 99.9% de los documentos estan en la categoría no relevante (Manning *et al*., 2009).

Si bien la UADY cuenta con un sistema bibliotecario SISBI que realiza búsquedas sobre una base de datos de trabajos de titulación, no es posible restringir esta búsqueda únicamente a la FCA, ya que agrupan los trabajos de titulación bajo el concepto de Ciencias Sociales y Humanidades, a la que pertenecen otras Dependencias de la Universidad, motivo por el cual se decidió implementar y hacer uso de DSpace (DuraSpace, 2002), para comparar su efectividad contra el sistema SIRETT. DSpace es un software de código abierto que gestiona repositorios de recursos digitales y es ampliamente conocido y utilizado en el ámbito de repositorios.

Se capturó en DSpace la misma información de los trabajos de titulación que se tenía en el sistema SIRETT (411 documentos) de modo que se pudiera validar el desempeño de dos sistemas de recuperación de información (algoritmos) diferentes sobre la misma base de datos. De esta manera, el objetivo fue comparar la efectividad de dos algoritmos de recuperación o máquinas de búsqueda mediante el cálculo de la precisión en ranking p, utilizando p = 10, es decir, los primeros 10 resultados de una búsqueda para ambos sistemas.

La hipótesis para evaluar la precisión es H_0: No existe diferencia significativa entre el sistema de recuperación SIRETT y el de DSpace.

Se solicitó a los profesores que realizaran cinco búsquedas en ambos sistemas (SIRETT y DSpace) sobre algún tema de su interés, resaltando el hecho de que las búsquedas deberían ser acordes a temas sobre trabajos de titulación de los alumnos de la Facultad de Ciencias Antropológicas entre 2011 y 2016. Para cada una de las cinco búsquedas, se les indicó que tomaran en cuenta solo los diez primeros resultados de cada sistema y que anotaran las posiciones de los trabajos de titulación que consideraran relevantes a la consulta que realizaron.

Una vez finalizadas las búsquedas, se calculó la precisión (3) para cada una de ellas para ambos rankings para cada uno de los documentos en donde estos fueron relevantes.

$$Precisión = \frac{\#\, relevantes\ (seleccionados)}{\#\, recuperados (10\ resultados)} \tag{3}$$

Después se calculó la precisión media del ranking de cada búsqueda para ambos sistemas, promediando los resultados de las precisiones de cada documento relevante, para finalmente utilizar la medida MAP (Mean Average Precision) que proporciona un resumen muy conciso de la eficacia de un algoritmo de recuperación durante muchas consultas (Manning, Raghavan *et al*., 2009). Así, por cada profesor se obtuvieron dos medidas MAP, una por cada motor de búsqueda (algoritmo). Estas muestras relacionadas son las que sirvieron para comparar el desempeño del motor del sistema de búsqueda SIRETT contra el de DSpace.

En la comparación del desempeño (MAP) entre sistemas se utilizó la prueba t de Student cuando las muestras son relacionadas (Wackerly *et al*., 2010); el paquete estadístico empleado fue el STATGRAPHICS Centurion XVII v. 17.0.16 (Statpoint, 2014) y la prueba se consideró significativa si $P < 0.05$.

En la comparación entre sistemas resultó que el desempeño del motor de búsqueda SIRETT no difirió significativamente ($t = 1.9125$, $P = 0.0719$) del desempeño de un gestor de repositorio de recursos digitales ampliamente utilizado en todo el mundo (DSpace). No se rechazó H_0, en otras palabras, el desempeño del motor de búsqueda de SIRETT se comportó de la misma manera que el de DSpace, con lo que se puede afirmar que SIRETT es capaz de recuperar satisfactoriamente los trabajos de titulación que los usuarios soliciten en sus consultas.

Una vez determinado que la efectividad de ambos sistemas es la misma, se procedió a analizar con la segunda parte de la evaluación, qué sucede con la percepción de ambos sistemas, es decir, cuál de ellos es el que los usuarios prefieren utilizar.

Utilidad, facilidad y percepción en la búsqueda y recuperación

Como se mencionó anteriormente, se decidió utilizar DSpace debido a no contar con homogeneidad en las opciones de búsqueda y recuperación entre SIRETT y

SISBI para evaluar la precisión en el resultado de sus búsquedas; sin embargo, el objetivo de la segunda parte de la evaluación de los profesores fue determinar la facilidad de uso entre ambos sistemas (SIRETT y SISBI) con el fin de realizar una comparación.

La hipótesis para evaluar la preferencia de uso es H_0: No existe diferencia significativa entre SISBI y SIRETT, respecto a la facilidad de uso percibida por los profesores.

Se les indicó a los profesores que realizaran cuatro actividades en ambos sistemas:

- Localizar por medio del título del documento, dos trabajos de titulación en los cuales haya participado como asesor o sinodal entre 2011 y 2016.
- Localizar por medio del autor del documento, algún trabajo de titulación en el cual haya participado como asesor o sinodal entre 2011 y 2016.
- Localizar cuántos trabajos de titulación existen en el año 2011 en su licenciatura.
- Escribir cuántos de los trabajos encontrados en la pregunta anterior son tesis.

En primer lugar, se aleatorizó el orden en que utilizarían cada sistema. Después se les indicó que trabajaran en el SIRETT en su apartado de búsqueda por metadatos y en el SISBI en su búsqueda avanzada, restringiendo los resultados a la biblioteca de Ciencias Sociales y Humanidades, al catálogo de trabajos de titulación y al rango de años entre 2011 y 2016.

Los profesores contaron con 30 minutos para realizar las cuatro actividades en ambos sistemas. Finalmente, se les solicitó que contestaran dos cuestionarios para validar la utilidad, facilidad y percepción de los sistemas.

Una de las técnicas ampliamente utilizadas para evaluar la usabilidad percibida por los usuarios de un sistema es emplear un cuestionario que debe completarse al finalizar la interacción con el producto, permitiendo tener una medición global que puede ser usada para comparar el resultado con sistemas similares.

La Escala de Usabilidad de un Sistema (SUS, System Usability Scale) (Brooke, 1996) es una herramienta muy utilizada para evaluar la usabilidad percibida por los usuarios luego de la interacción con un sistema (Tullis y Albert, 2013). Se trata de un cuestionario con 10 ítems que los usuarios puntúan según su nivel de aceptación, utilizando una escala Likert de 1 (muy en desacuerdo) a 5 (muy de acuerdo). La mitad de los ítems se expresan positivamente y la otra mitad negativamente. Se utiliza un algoritmo para obtener un puntaje total de 0 a 100, donde 100 representa el mejor puntaje en términos de una visión global de la percepción que tuvo un usuario sobre la usabilidad del sistema.

SUS ha sido utilizado ampliamente en estudios de usabilidad de sistemas y aplicaciones tanto en la industria como en la academia para comparar sistemas (Tullis y Albert, 2013), lo que garantiza su efectividad como herramienta para medir la usabilidad percibida de un sistema.

Como resultado de la evaluación, se obtuvo 61.97% para SISBI y 88.82% para SIRETT, con lo cual puede concluirse que los profesores encuentran a SIRETT más fácil de usar que SISBI.

El segundo de los cuestionarios utilizados se denomina Modelo de Aceptación de Tecnología (TAM, Technology Acceptance Model) (Davis, 1989) que mide la intención de uso de una tecnología en términos de la utilidad y facilidad de uso percibida por el usuario. El instrumento ha sido analizado en términos de su robustez y validez, se ha empleado en numerosos estudios a lo largo del tiempo (Marangunić & Granić, 2015; Lee *et al*., 2003). TAM provee una escala simple de 12 preguntas que dan una visión general de la percepción del usuario, midiendo tres aspectos: la utilidad (UT), la facilidad (FA) y la percepción (PE) (Tabla 2).

El cuestionario TAM consta de seis reactivos (R_i i=1,2,..,6) que se refieren a la UT y otros seis (R_i i=7,…,12) a la FA, de modo que la PE la conforman los doce reactivos. Dichos reactivos tienen una escala de Likert de 7 puntos cuyos valores cubren el rango del valor 1 (“totalmente en desacuerdo”) hasta el 7 (“totalmente de acuerdo”).

Sea CR_i la calificación dada por el usuario al i-ésimo reactivo. Así, la utilidad (4) para un usuario se obtuvo de la siguiente forma

$$UT = \frac{\frac{1}{6}\sum_{i=1}^{6} CR_i}{7} = \frac{\sum_{i=1}^{6} CR_i}{42} \tag{4}$$

Dado que $6 \leq \sum_{i=1}^{6} CR_i \leq 42$ entonces se tiene

$$\frac{1}{7} \leq \frac{\sum_{i=1}^{6} CR_i}{42} \leq 1 \tag{5}$$

Por lo que la utilidad para un usuario toma valores entre $0.1429 \leq UT \leq 1$. Análogamente sucede con la facilidad. Y la percepción está dada por la siguiente fórmula (6)

Tabla 2. Preguntas del modelo de aceptación tecnológica

Utilidad
Pregunta:
1. Usar la herramienta me permite realizar las tareas con mayor rapidez?
2. Usar la herramienta mejora mi desempeño?
3. Usar la herramienta facilita la realización de mis actividades de trabajo?
4. Usar la herramienta mejora mi eficacia en el trabajo?
5. Interactuar con la interfaz de la herramienta aumenta mi productividad?
6. La herramienta me resulta útil en mi trabajo?
Facilidad de uso
Pregunta:
7. Me resulta fácil que la herramienta haga lo que quiero que realice?
8. Mi interacción con la herramienta es clara y entendible?
9. Aprender a utilizar la herramienta me resultó fácil?
10. Me resultó sencillo adquirir destreza en el uso de la herramienta?
11. Encuentro la herramienta fácil de utilizar?
12. Considero que la herramienta es flexible para interactuar con ella?

$$PE = \frac{\frac{1}{12}\sum_{i=1}^{12} CR_i}{7} = \frac{\sum_{i=1}^{12} CR_i}{84} \quad (6)$$

Procediendo análogamente se tiene que $0.1429 \leq PE \leq 1$. Se utilizó la prueba t de Student cuando las muestras son relacionadas (Wackerly *et al*., 2010) para comparar la facilidad entre herramientas. Debido al incumplimiento del supuesto de normalidad, se aplicó la prueba de rangos con signo de Wilcoxon (Wackerly *et al*., 2010) para comparar entre herramientas la utilidad y la percepción. Las pruebas estadísticas se consideraron significativas cuando $P < 0.05$, y se utilizaron los paquetes estadísticos STATGRAPHICS Centurion XVII v. 17.0.16 (Statpoint, 2014) y el SPSS 22 (IBM, 2013).

Entre las herramientas SIRETT y SISBI difirió la utilidad (T=5.50, P=0.0004), la facilidad (t=3.6470, P=0.0018, gl=18) y la percepción (T=8.50, P=0.0005); resultando significativamente mayor la utilidad, facilidad y la percepción en SIRETT que en SISBI (Tabla 3). Con base en los resultados se rechaza H_0 y se concluye que existe mayor aceptación de SIRETT sobre SISBI.

Pruebas del sistema de recomendación híbrido con alumnos

Para efectos de estas pruebas, se contó con la participación de dos grupos de estudiantes de la Facultad de Ciencias Antropológicas, el primer grupo lo conformaron 32 alumnos del quinto semestre de la Licenciatura en Comunicación Social que se denominó "tesistas". La asignatura que cursaron se llam*ó Métodos y Técnicas de Investigación en Comunicación*.

El segundo grupo de alumnos lo conformaron 42 estudiantes de la Licenciatura en Turismo de tercer semestre a los que se denominó "alumnos general" y la asignatura que cursaron fue Fundamentos de Investigación.

Tabla 3. Usabilidad entre las herramientas SIRETT y SISBI por profesores (n=19)

Variable	Herramienta	*Promedio	DE	Mediana	*Suma de rangos de d=SIRETT-SISBI	Mínimo	Máximo
UT	SIRETT	0.8709	0.1429	0.9048	T+ =130.50 a	0.5	1
	SISBI	0.6905	0.2302	0.7143	T- =5.50 b	0.2857	1
FA	SIRETT	0.9236 a	0.0850	0.9286	----------	0.7381	1
	SISBI	0.7068 b	0.2416	0.7143		0.2143	1
PE	SIRETT	0.8972	0.1033	0.9167	T+ =144.50 a	0.6191	1
	SISBI	0.6986	0.2136	0.7262	T- =8.50 b	0.3333	1

** Promedios o suma de rangos con distinta letra difieren ($P<0.05$), prueba t o Wilcoxon. T^+ y T^-: suma de rangos de las diferencias positivas y negativas*

Rendimiento de la recomendación por uso

Para evaluar la recomendación por uso se realizó un estudio en el que participaron los alumnos tesistas. El objetivo fue contrastar la utilidad del uso del filtrado de palabras clave correspondientes al perfil del usuario, el cual se conforma de las palabras clave que ha utilizado históricamente en sus consultas al sistema.

Se utilizó como medida la tasa media de éxito recíproco (7) también conocida como Mean Reciprocal Rank (MBR) de cada consulta, que es el recíproco del rango o posición del primer elemento seleccionado en la lista o ranking recomendado o cero en el caso de que ningún elemento haya sido seleccionado. La puntuación correspondiente para un conjunto de consultas es la media de las posiciones recíprocas de cada una de las consultas (Bian *et al*., 2008).

$$MRR = \frac{1}{|B|}\sum_{i=1}^{|B|}\frac{1}{rank_i} \qquad (7)$$

Donde

B = número de búsquedas
$rank_i$ = rango o posición del primer elemento seleccionado para la búsqueda i

Los valores más altos de MRR indican que la lista o ranking correspondiente contiene los elementos seleccionados en las posiciones más altas o primeras posiciones de la lista y, por tanto, para nuestra propuesta de recomendación estos valores más altos de MRR muestran un mejor ranking al utilizar el filtrado de palabras clave.

Hay dos acercamientos para medir la exactitud de un ranking. Podemos tratar de determinar el orden correcto de una lista de items para un usuario y medir qué tan cerca llega el sistema a este orden correcto o podemos intentar medir la utilidad del ranking del sistema para un usuario. Debido a la naturaleza del estudio donde se le preguntaba al usuario su selección del listado de los 10 primeros resultados y cuáles le resultaban interesantes de arriba hacia abajo, se decidió utilizar la métrica Mean Reciprocal Rank bajo el supuesto de que, en primera era un estudio presencial (online) y en segunda, la utilidad de la lista de recomendación era acumulativa, dada por la suma de las utilidades de las recomendaciones individuales. No se consideró la métrica Normalized Distance-based Performance Measure por el hecho de no contar con un ranking de referencia y estar enfocado a estudios no presenciales (offline) (Shani y Gunawardana, 2011).

La hipótesis para evaluar el rendimiento de la recomendación por uso es H_0: No existe diferencia significativa en el uso del filtrado por palabras clave en la recomendación por uso.

Se les solicitó a los estudiantes que consultaran (abrieran en el navegador) cinco trabajos de titulación para cada uno de cinco temas establecidos (redes sociales, educación, racismo, tradiciones y violencia). Podrían utilizar como criterios de búsqueda lo que consideraran relacionado al tema principal y su búsqueda podría ser tan general o específica como ellos determinaran.

Una vez que los estudiantes terminaron de realizar las búsquedas se les indicó que se dirigieran al apartado de recomendación por uso, donde encontrarían un listado que proporciona el sistema por defecto. Realizando una revisión de los títulos de los trabajos de titulación presentados en el listado o ranking de arriba hacia abajo, se les solicitó que localizaran solo uno que consideraran el representativo para cada uno de los cinco temas establecidos al inicio de la prueba, sin hacer caso de las palabras clave mostradas al inicio del ranking, después se les solicitó que escribieran en una hoja la posición correspondiente al documento o cero en caso de no encontrar ninguno.

Posteriormente se les indicó que hicieran uso del filtrado de palabras clave seleccionado solo aquellas palabras clave que tuvieran relación con el primer tema (redes sociales) y en este nuevo listado reordenado localizaran el trabajo de titulación que anteriormente habían considerado representativo del mismo tema y escribieran su nueva posición o cero en caso de no encontrarlo. Lo mismo se hizo para los otros cuatro temas restantes. Finalmente, se les solicitó que contestaran los cuestionarios SUS y TAM.

A modo de ejemplo, se presenta la Tabla 4 donde se muestran las posiciones en el ordenamiento de los trabajos de titulación recomendados por uso para el usuario con id=398, sin utilizar el filtrado por palabras clave y utilizando el filtrado por palabras clave para cada uno de los cinco temas establecidos.

Tabla 4. Posiciones de los trabajos de titulación representativos de cada tema, seleccionados por el usuario con id=398, sf = sin filtrado, cf = con filtrado, T = tema

T1		T2		T3		T4		T5	
sf	cf	sf	cf	sf	cf	sf	cf	sf	cf
8	5	11	2	3	3	27	4	10	2

Se puede observar que cuando se utiliza el filtrado por palabras clave sobre el ranking de la recomendación

por uso que proporciona el sistema por defecto, se logran posicionar los documentos en las posiciones más altas del listado.

Con los resultados de las posiciones con y sin filtro de cada uno de los participantes y utilizando la métrica MRR, se procedió a comparar el rendimiento de SIRETT. Se utilizó la prueba t de Student cuando las muestras son relacionadas (Wackerly *et al.*, 2010) para comparar el rendimiento (MRR) de SIRETT sin filtrado (sf) y con filtrado (cf). Las prueba estadística se consideró significativa cuando $P < 0.05$, y se utilizó el paquete estadístico STATGRAPHICS Centurion XVII v. 17.0.16 (Statpoint, 2014).

El rendimiento de SIRETT con y sin filtrado difirió significativamente ($t = -7.4284$, $P<0.0001$, $gl=31$), resultando mayor la MRR con filtrado (Tabla 5). Con base en los resultados se rechaza H_0 y se concluye un mayor rendimiento de la recomendación por uso al utilizar el filtrado por palabras clave.

Usabilidad

El objetivo de las pruebas que se realizaron con el segundo grupo de estudiantes (alumnos general) fue validar la usabilidad de SIRETT. Para tal efecto se les mostró el funcionamiento completo del sistema, desde sus apartados de búsqueda hasta la recomendación y se les pidió que contestaran los dos cuestionarios utilizados anteriormente con los profesores (SUS y TAM).

La hipótesis para evaluar la preferencia de uso por medio de SUS es H_0: El sistema SIRETT es fácil de usar por parte de alumnos tesistas y alumnos general.

Como resultado de la prueba SUS para alumnos tesistas se obtuvo una usabilidad de 83.33%. Del mismo modo, se obtuvo para los alumnos en general una usabilidad con la misma prueba de 82.56%. Estos resultados indican una aceptación positiva por parte de ambos tipos de usuarios, por lo tanto, no se rechazó H_0.

Comparación de la utilidad, facilidad y percepción entre los tres tipos de usuarios

Como se mencionó al principio del artículo, los tres tipos de usuarios (profesores, tesistas y alumnos general) contestaron el cuestionario TAM, debido a esto, se procedió al análisis de los resultados estadísticamente con el fin de determinar si hay diferencias entre la tres variables que maneja la prueba (utilidad = UT, facilidad = FA y percepción = PE).

La hipótesis para evaluar la preferencia de uso por medio de TAM es H_0: No existe diferencia significativa en la utilidad, facilidad y percepción entre los tres tipos de usuarios en el uso del sistema SIRETT.

Debido al incumplimiento del supuesto de normalidad se utilizó la prueba de Kruskal-Wallis (Wayne, 1990) para comparar entre tipos de usuario la utilidad, facilidad y la percepción acerca de SIRETT. Las pruebas estadísticas se consideraron significativas cuando $P < 0.05$, y se utilizó el paquete estadístico STATGRAPHICS Centurion XVII v. 17.0.16 (Statpoint, 2014).

En la comparación de SIRETT entre tipos de usuario no se obtuvo diferencias significativas respecto a la: utilidad ($H=0.7948$, $P=0.6721$, $gl=2$; Tabla 6), facilidad ($H=2.0990$, $P=0.3501$, $gl=2$; Tabla 7) y percepción ($H=2.1348$, $P=0.3439$, $gl=2$; Tabla 8), por lo tanto, no se rechazó H_0. Cabe señalar que los posibles valores de UT, FA y PE se encuentran entre 0.1429 y 1, y en cada caso los promedios, medianas, mínimos y máximos indican valores altos para la utilidad, facilidad y la percepción, esto también puede notarse en las distribuciones de los valores (Figuras 3, 4, 5).

Análisis del desempeño de la interfaz Web

El desempeño de la interfaz Web de SIRETT fue validado utilizando Yahoo! YSlow (Yahoo, 2015), una herramienta que analiza una página Web y le otorga un grado según una colección de reglas.

Vences *et al.* (2016) demuestran como combinando distintas configuraciones al servidor Web Apache, como el manejo de caché y la compresión de datos, así como siguiendo buenas prácticas al utilizar las sentencias SQL en el manejador de base de datos, se obtiene una reducción significativa (poco más de 50%) en los tiempos de respuesta del sistema SIRETT. Se observa también una reducción en el volumen de datos transferidos (reducción de 90.8%) entre el servidor y el cliente, lo cual tiene una repercusión inmediata en la percepción del usuario, ya que la página se despliega más rápido. Por otro lado, se demuestra que para los casos

Tabla 5. Rendimiento de SIRETT con (cf) y sin filtrado (sf). (n=32 alumnos tesistas)

Métrica	Condición	*Promedio	DE	Mediana	Mínimo	Máximo
MRR	sf	0.2095 a	0.0947	0.1872	0.0473	0.4034
	cf	0.3851 b	0.1546	0.3811	0.1139	0.6307

** Promedios con distinta letra difieren (P<0.05), prueba t*

Tabla 6. Utilidad de SIRETT entre tipos de usuario

Tipo de usuario	n	Promedio	DE	Rango promedio*	Mediana	Mínimo	Máximo
Profesor (P)	19	0.8709	0.1429	56.9474 a	0.9048	0.5	1
Alumno general (AG)	42	0.8532	0.1408	52.1190 a	0.8690	0.5	1
Alumno tesista (AT)	42	0.8424	0.1362	49.6429 a	0.8571	0.5	1

** Rangos promedio con igual letra no difieren (P>0.05), prueba de Kruskal-Wallis*

Tabla 7. Facilidad de SIRETT entre tipos de usuario

Tipo de usuario	n	Promedio	DE	Rango promedio*	Mediana	Mínimo	Máximo
Profesor (P)	19	0.9236	0.0849	57.8684 a	0.9286	0.7381	1
Alumno general (AG)	42	0.9059	0.1147	54.1667 a	0.9524	0.5238	1
Alumno tesista (AT)	42	0.8957	0.0923	47.1786 a	0.9048	0.5952	1

** Rangos promedio con igual letra no difieren (P>0.05), prueba de Kruskal-Wallis*

Tabla 8. Percepción de SIRETT entre tipos de usuario

Tipo de usuario	n	Promedio	DE	Rango promedio*	Mediana	Mínimo	Máximo
Profesor (P)	19	0.8972	0.1033	58.6316 a	0.9167	0.6190	1
Alumno general (AG)	42	0.8795	0.1176	53.7262 a	0.8988	0.5476	1
Alumno tesista (AT)	42	0.8690	0.0953	47.2738 a	0.8809	0.6309	1

** Rangos promedio con igual letra no difieren (P>0.05), prueba de Kruskal-Wallis*

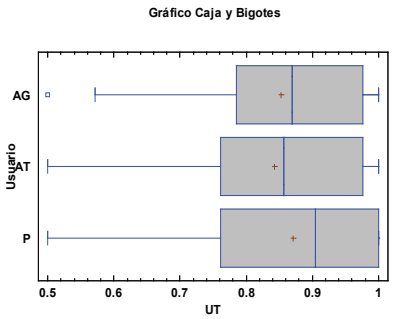

Figura 3. Distribución de la utilidad entre tipos de usuario

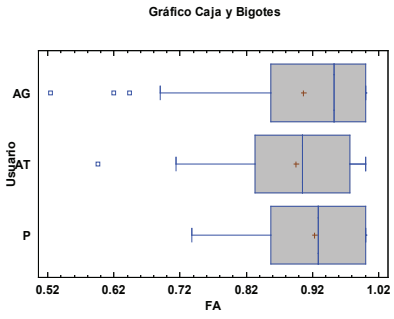

Figura 4. Distribución de la facilidad entre tipos de usuario

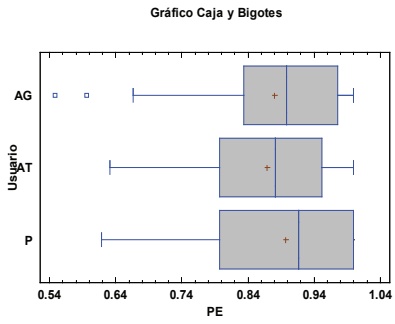

Figura 5. Distribución de la percepción entre tipos de usuario

donde la interacción del sistema con la base datos es primordial y se requiere un gran uso de la misma, como la creación de índices para un sistema de recuperación de información, y siguiendo las recomendaciones en la construcción y ejecución de sentencias INSERT, se logra una reducción significativa (reducción de 99.98%) en los accesos a la base de datos, lo que repercute directamente en una reducción en el tiempo de procesamiento.

Conclusiones

En este artículo se explica cómo evaluar un sistema de recomendación. Desde los criterios basados en los algoritmos de recomendación, hasta los criterios externos, como por ejemplo, la vista del usuario. Se utilizó como caso de estudio el SIRETT. Para ello se contó con la participación de profesores y alumnos de la FCA-UADY.

Los resultados obtenidos en los estudios muestran que la precisión del motor de búsqueda y recuperación de SIRETT es tan robusta como la de un sistema gestor de repositorios utilizado alrededor del mundo (DSpace), no encontrando diferencias significativas entre ambos algoritmos.

Se demostró también que existe diferencia significativa en la utilidad, facilidad y percepción entre SIRETT y el SISBI de la UADY, teniendo una mayor aceptación el primero sobre el segundo.

Las pruebas realizadas determinaron que el uso del filtro de palabras clave en el listado de recomendación resultó en posicionar en las primeras posiciones del ranking a los documentos que tenían mayor relación con la palabra clave filtrada según los propios alumnos.

En la comparación de la utilidad, facilidad y percepción del SIRETT entre los tres tipos de usuarios, no se obtuvo diferencias significativas. Cabe señalar que los posibles valores de UT, FA y PE se encuentran entre 0.1429 y 1, y en cada caso los promedios, medianas, mínimos y máximos resultaron en valores altos.

Respecto a la usabilidad de SIRETT, como resultado de la prueba SUS para profesores, alumnos tesistas y alumnos general se obtuvo 88.82%, 83.33% y 82.56%, respectivamente. Estos resultados nos indican una aceptación positiva por parte de los tres tipos de usuarios.

Finalmente, con base en los estudios realizados el SIRETT podría ser un SR opcional o complementario al SISBI.

Como una línea de investigación futura podría implementarse el uso de la información generada por los usuarios para analizarla con técnicas de minería de datos con el objetivo de identificar grupos basados en los trabajos de titulación consultados o de extraer reglas de asociación y clasificación con el objetivo de mejorar la recomendación.

Referencias

Alejandres, H., González, J., González, N. (2016). Efecto de explicaciones sobre la confianza del usuario en sistemas de recomendación sensibles al contexto. *Ingeniería Investigación y Tecnología,* 17(4) 419-428. Recuperado el 7 de agosto de 2017 de https://doi.org/10.1016/j.riit.2016.11.001.

Bian, J., Liu, Y., Agichtein, E., Zha, H. (2008). Finding the right facts in te crowd: factoid question answering over social media. En Proceedings of the 17th International conference on World Wide Web, 467-476. 2008. Recuperado el 7 de agosto de 2017 de http://dl.acm.org/citation.cfm?id=1367561.

Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A. (2013). Recommender systems survey. K*nowledge-Based Systems*, 46, 109-132. Recuperado el 7 de agosto de 2017 de http://doi.org/10.1016/j.knosys.2013.03.012.

Brooke, J. (1996). SUS - A quick and dirty usability scale. *Usability evaluation in industry*, 189 (194) 4-7. 1996. Recuperado el 7 de agosto de 2017 de https://www.researchgate.net/publication/228593520_SUS_A_quick_and_dirty_usability_scale.

Croft, B., Metzler, D., Strohmann, T. (2015). *Search Engines: Information Retrieval in Practice*. Pearson Education, Inc. Recuperado el 7 de agosto de 2017 de http://ciir.cs.umass.edu/downloads/SEIRiP.pdf.

Davis, F. (1989). Perceived usefulness, perceived ease of use and, user acceptance of information technology. *MIS Quarterly*, 13(3), 319. Recuperado el 7 de agosto de 2017 de https://pdfs.semanticscholar.org/3969/e582e68e418a2b-79c604cd35d5d81de9b35d.pdf.

DuraSpace. DSpace. 2002. Recuperado de http://dspace.org/

Erdt, M., Fernandez, A., Rensing, C. (2015). Evaluating recommender systems for technology enhanced learning: A quantitative survey. *IEEE Transactions on Learning Technologies*, 8(4). Recuperado el 7 de agosto de 2017 de http://doi.org/10.1109/TLT.2015.2438867.

Guy S. y Asela G. (2011). Recommender systems handbook en Ricci F., Rokach L., Shapira B., Kantor P. editores. Dordrecht Heidelberg London: Springer New York. Recuperado de http://doi.org/10.1007/978-0-387-85820-3.

IBM. (2013). *SPSS Statistics for Windows.* Armonk, NY: IBM Corp.

Lee, Y., Kozar, K.A., Larsen, K.R.T. (2003). The technology acceptance model: Past, Present, and Future. *Communications of the Association for Information Systems* 12(50). Recuperado de http://aisel.aisnet.org/cais/vol12/iss1/50

Manning, C., Raghavan, P., Schütze, H. (2009). *An introduction to information retrieval.* Cambridge, England: Cambridge University Press. Recuperado el 7 de agosto de 2017 de https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf.

Marangunić, N. y Granić, A. (2015). Technology acceptance model: a literature review from 1986 to 2013. *Universal Access in the Information Society*, 14(1), 81-95. Recuperado de https://doi.org/10.1007/s10209-014-0348-1.

Pressman, R. (2010). *Ingeniería del software, un enfoque práctico*. México: McGraw Hill. Recuperado el 7 de agosto de 2017 de http://artemisa.unicauca.edu.co/~cardila/Libro_Pressman_7.pdf.

Ricci, F., Rokach, L., Shapira, B., Kantor, P. (2011). *Recommender systems handbook*. USA: Springer New York Dordrecht Heidelberg London. Recuperado el 7 de agosto de 2017 de http://doi.org/10.1007/978-0-387-85820-3.

Statpoint, I. (2014). STATGRAPHICS Centurion XVII. Recuperado de www.statgraphics.com

Tullis, T. y Albert, W. (2016). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics*. Amsterdam: Morgan Kaufmann.

Van, C. (1979). Information retrieval. University of Glasgow, 4-7. Recuperado el 7 de agosto de 2017 de http://openlib.org/

home/krichel/courses/lis618/readings/rijsbergen79_infor_retriev.pdf.

Vences, R., Menéndez, V.H., Zapata A. (2015). Sistema de recomendación para la búsqueda personalizada en un repositorio de trabajos de titulación. *Revista Latinoamericana de Ingeniería de Software (ReLAIS)*. 3(6), 223-230. Recuperada el 7de agosto de 2017 de https://doi.org/10.18294/relais.2015.223-230.

Vences R., Menéndez V.H., Zapata A. (2016). Optimización del desempeño de un sistema de recomendación de documentos de texto basado en la configuración de los servidores. *Revista electrónica de Computación, Informática, Biomédica y Electrónica (RECIBE)*, 5 (2). ISSN: 2007-5448. Recuperada el 7 de agosto de 2017 de http://www.revistascientificas.udg.mx/index.php/REC/article/view/5781/5297.

Wackerly, D.D., Mendenhall, W., Scheaffer, R.L. (2010). *Estadística matemática con aplicaciones*, 7a ed. CENGAGE Learning.

Wayne, W.D. *Applied nonparametric statistics*. (1990). 2a ed., USA: D. P. G. CA, Ed.

Wu W., He L., Yang J. (2012). Evaluating recommender systems. En 7th International Conference on Digital Information Management, 56-61. Recuperado el 7 de agosto de 2017 de http://doi.org/10.1109/ICDIM.2012.6360092.

Yahoo Y.Slow. (2015). Recuperado el 7 de agosto de 2017 de http://developer.yahoo.com/yslow/