International Journal of Information Systems and Software Engineering for Big Companies (IJISEBC)

Recibido: 02-09-2016 / Revisado: 24-11-2016
Aceptado: 24-11-2016 / Publicado: 15-12-2016

Una comprensiva revisión de los métodos de recomendación basados en técnicas probabilísticas

A comprehensive view of recommendation methods based on probabilistic techniques

Priscila Valdiviezo-Díaz[1], Antonio Hernando[2]

[1] Universidad Técnica Particular de Loja, Ecuador
[2] Universidad Politécnica de Madrid, España

pmvaldiviezo@utpl.edu.ec , antonio.hernando@upm.es

RESUMEN. Esta investigación tiene como objetivo utilizar un método de recomendación híbrido basado en técnicas probabilísticas y de modelado de tópicos que brinde al usuario recomendaciones más ajustadas frente a los modelos de recomendación tradicionales. Este artículo presenta una revisión comprensiva de los métodos de recomendación para sistemas basado en contenido y filtrado colaborativo. Entre los métodos analizados están las Matrices de Factorización Probabilística y el método de Asignación Latente de Dirichlet. La revisión de la literatura entorno a estos modelos se centra en la identificación de problemas y cuestiones abiertas que pueden ser abarcadas para futuras investigaciones. Se analiza el funcionamiento de algunos modelos de recomendación que integran técnicas de factores latentes y de modelado de tópicos, que serán de base para comparar los resultados obtenidos con el modelo híbrido.

ABSTRACT. This research aims to use a hybrid recommendation method based on probabilistic techniques and topics modeling that provide recommendations most close fitting the user compared to other traditional recommendation models. We carry out a comprehensive review of the recommended methods for content-based systems and collaborative filtering, mainly in the domain of recommending movies. The methods discussed are the matrix factorization and Latent Dirichlet Allocation method. The literature review around these models focuses on identifying problems and open issues that may be covered for future researches. Also, we analyzed the recommendation models that integrant latent factor methods and topics modeling, which will be used to compare results obtained with the hybrid model.

PALABRAS CLAVE: Sistema recomendador, Matriz de factorización, Modelo de tópicos, Asignación latente de Dirichlet.

KEYWORDS: Recommender system, Factorization matrix, Topic model, Latent Dirichlet Allocation.

Valdiviezo-Díaz, P.; Hernando, A. (2016). Una comprensiva revisión de los métodos de recomendación basados en técnicas probabilísticas. *International Journal of Information Systems and Software Engineering for Big Companies (IJISEBC), 3*(2), 65-74.

1. Introducción

El gran volumen de información disponible en Internet hace que los usuarios inviertan gran parte de su tiempo en buscar y obtener información relevante de esta red. En general cada vez hay más productos y servicios disponibles en Internet que son de interés de los usuarios con gustos y preferencias diferentes. De ahí que, los sistemas de recomendación (SR) se convierten en una alternativa para ayudar a los usuarios a obtener información precisa y personalizada de grandes repositorios de información que pueden estar disponibles en Internet.

Los sistemas de recomendación son ampliamente utilizados en la Web para recomendar productos y servicios a los usuarios (Bing, 2011). En estos sistemas, recomendaciones personalizadas sobre los ítems son generados a través de la predicción de las preferencias de usuarios (Huang, Chung y Hsinchun, 2004), basados en el análisis de las preferencias pasadas o en las preferencias de usuarios similares.

Algunas técnicas de filtrado de información se pueden utilizar en los sistemas de recomendación, entre ellas, las más conocidas son: filtrado basado en contenido (CBF) y el filtrado colaborativo (CF) (Ricci, Rokach, Shapira y Kantor, 2011). En estos sistemas de recomendación, los comportamientos de los usuarios se ven influenciados por los intereses ocultos de los usuarios (Liu, Chen, Member, Xiong, Ding y Chen, 2012), información que es muy importante conocer para proporcionar mejores recomendaciones.

En la literatura se puede encontrar algunos trabajos relacionados a los Sistemas Recomendadores aplicados a una variedad de ámbitos como el comercio electrónico, educación, salud, etc., donde es necesario el uso de ciertas técnicas y métodos para el proceso de recomendación, entre ellos están los métodos de Matriz de Factorización (Koren, Bell y Volinsky, 2009) que actualmente están recibiendo mayor atención principalmente en la descomposición de variables latentes. El método Latent Dirichlet Allocation (LDA) (Blei, Ng y Jordan, 2003) también es otro modelo latente que puede ser utilizado para encontrar contenido semántico oculto en un corpus de texto. Ambos modelos actualmentes han sido extendidos para ser utilizados en sistemas recomendadores para modelar las preferencias del usuario como factores latentes.

Con el fin de clarificar todas las cuestiones involucradas en los sitemas recomendadores basados en métodos probabilísticos, en las siguientes secciones se presenta el estado del arte realizado en este trabajo, enfocándonos en los métodos de recomendación para sistemas basado en contenido y filtrado colaborativo. Entre los métodos analizados están las Matrices de Factorización Probabilística (PMF, por sus siglas en inglés) y el método de Asignación Latente de Dirichlet (LDA, por sus siglas en inglés) para el modelado del contenido de los items como factores latentes. Como parte de la revisión de la literatura se presenta también algunos de los problemas encontrados y las cuestiones abiertas que pueden ser abarcadas para futuras investigaciones. Seguidamente, se presenta una breve descripción del análisis del funcionamiento de algunos de los algoritmos que van a ser utilizados en esta investigación. Finalmente, se presentan las conclusiones de este trabajo.

2. Estado del arte

En este apartado se presentan los enfoques de recomendación mayormente utilizados en sistemas recomendadores y métodos probabilísticos relacionados con el modelado de factores latentes, como aquellos basados en matrices de factorización (Koren, Bell y Volinsky, 2009) y el modelado de tópicos (Hofmann, 1999). En nuestro caso el principal enfoque está en los modelos de matriz de factorización probabilística (Salakhutdinov y Mnih, 2008) y los métodos de Asignación Latente de Dirichlet (Blei, Ng y Jordan, 2003) que trabajan sobre documentos, puesto que la idea es utilizar un modelo que combine ambas técnicas.

A. Enfoques de Recomendación

En la literatura se puede encontrar diferentes enfoques de sistemas recomendadores, en este caso presentamos los más conocidos, por ejemplo en (Ghauth y Abdullah, 2010) se mencionan:

Sistemas basados en contenido: que pueden ser diseñados para recomendar items similares a los que al usuario le han gustado en el pasado (Lops, De Gemmis y Semeraro, 2011). En este tipo de filtrado el contenido de los ítems es importante para predecir su relevancia basado en un perfil de usuario, el cual incluye los gustos, preferencias y otras características, y sólo los ítems que tienen un alto grado de similitud con el perfil del usuario son recomendados (Chang y Hsiao, 2013). En este sentido el top-N de los mejores items o items más similares se recomiendan al usuario.

Sistemas de filtrado colaborativo: Estos sistemas basan sus predicciones y recomendaciones en las calificaciones o comportamiento de otros usuarios en el sistema (Ekstrand, Riedl y Konstan, 2010). Buscan similaridad entre usuarios y hacen sugerencias de ítems que fueron considerados por otros usuarios en el pasado, en otras palabras, identifica ítems basándose en las opiniones de usuarios "similares" al que se le va hacer la recomendación, y tiene la capacidad de formar grupos de usuarios afines a un mismo ítem. Generalmente las técnicas de filtrado colaborativo se agrupan dentro de dos categorías: basada en memoria y basada en modelos. Los sistemas recomendadores basados en memoria usan el método KNN (K-Nearest-Neighbour, por sus siglas en inglés), para predecir los ratings que los usuarios darían a los ítems (Hernando, Jesus y Fernando, 2016; Wen, 2008). Los sistemas recomendadores basados en modelos, utilizan un modelo para predecir los ratings que hacen los usuarios. Los algoritmos de esta categoría incluyen enfoques como matrices de factorización (Ahamed y Parambath, 2013), basados en grafos (Huang, Chung y Hsinchun, 2004; Jin Rong y Cheng Xiang, 2003) y otros modelos de factores latentes.

Sistemas híbridos: Combinan dos o más enfoques de recomendación para tener un mejor funcionamiento. Se utilizan comúnmente el filtrado colaborativo con otra técnica que reduzca problemas de recomendación con nuevos ítems. Estos sistemas tratan de mejorar todas las limitaciones que tienen los demás tipos de sistemas recomendadores.

B. Matrices de factorización

Son métodos de filtrado colaborativo basado en modelos (Ahamed y Parambath, 2013), donde la idea general es modelar las interacciones usuario-item como factores que representan características latentes de los usuarios y los elementos del sistema.

En (Benítez, Escudero, Kanaan y Masip, 2014) los métodos de factorización matricial consisten en descomponer en factores una matriz de datos. Esta técnica juega un importante rol en diversos campos, por ejemplo, en sistemas de recomendación, los modelos de matriz de factorización mapean usuarios e ítems en un espacio de factores latente de dimensionalidad f, donde una alta correspondencia entre los factores de los usuarios y los ítems conduce a una recomendación (Ricci, Rokach, Shapira y Kantor, 2011).

A fin de evitar el sobre ajuste (overfitting), se debe regularizar los parámetros aprendidos, cuyas magnitudes son penalizadas. En (Salakhutdinov y Mnih, 2008) se ofrece una base probalísitica para la regularización, que se presenta como una factorización probabilística de matrices. El modelo de matriz de factorización probabilistica es un simple modelo de análisis factorial, que consiste en un modelo lineal gaussiano de variable latente restringido.

En este modelo se define la distribución condicional sobre las calificaciones observadas, Rij que representa el rating del item i por el usuario u, dada por las matrices latentes U y V (Blei, Ng y Jordan, 2003; Lops, De Gemmis y Semeraro, 2011). La predicción del rating se lleva a cabo multiplicando el vector de usuario correspondiente y vector de ítems Uu y Vi respectivamente.

C. Método de modelado de tópicos: LDA

La Asignación Latente de Dirichlet (LDA) es un modelo generativo probabilístico no supervisado para modelar grandes corpus de texto, y generar aleatoriamente los documentos que se observan en este corpus

(Blei, Ng y Jordan, 2003). En este modelo se asume que existen varios temas para un corpus y un documento puede ser tomado como una bolsa de palabras (wi, j) que son generados por estos temas (Yu, Zhang y Zhu, 2012). Donde, cada documento es modelado como una mezcla aleatoria sobre tópicos que son latentes, y cada tópico se caracteriza como una distribución de probabilidad sobre las palabras, es decir, distribuciones multinomiales que consisten en dar a cada palabra del vocabulario una probabilidad, donde las palabras con alta probabilidad están más asociadas con ese tópico que las palabras con baja probabilidad (Blei, Ng y Jordan, 2003; Richert y Coelho, 2013). Los tópicos son considerados como grupos (clústeres) de palabras que vienen junto con cierta probabilidad, en este caso cada clúster es un tópico.

En trabajos como (Chang y Hsiao, 2013) se ha demostrado que LDA es capaz de capturar la información semántica latente de una colección de documentos, superior en comparación con varios otros modelos. En sistemas recomendadores LDA ha sido utilizado para el análisis del contexto en métodos basados en contenido (Yu, Zhang y Zhu, 2012). En (Ekstrand, Riedl y Konstan, 2010), los usuarios son representados por sus factores latentes (como una distribución $P(k \mid u)$, que son instancias de una variable aleatoria extraída de una distribución de Dirichlet, donde k es una variable latente que pertenece al conjunto de tópicos latentes K. Este modelo requiere de dos hyper-parámetros que se pueden aprender, α y β, donde α hace referencia a un vector de parametrización de la distribución de Dirichlet, a partir de la cual se extraen los usuarios, y β, es representada por una matriz de probabilidad de tópicos – ítems, representada por $P(i|k)$. LDA también puede ser utilizado en sistemas recomendadores basados en filtrado colaborativo, donde se asume que los documentos y las palabras en el documento son análogos a los usuarios e ítems respectivamente (Blei, Ng y Jordan, 2003) y los tópicos llegan a ser los intereses ocultos (Liu, Chen, Member, Xiong, Ding y Chen, 2012). Además, como se observa en la figura 1, el número de ocurrencias de cada palabra en el documento, se puede asumir como el rating que el usuario ha dado a cada item.

Figura 1. Analogía entre los elementos del método LDA basado en contenido frente al de filtrado colaborativo.

En este caso, los tópicos latentes son asumidos para ser distribuidos multinomialmente sobre los usuarios, y los ítems del usuario se supone que se distribuyen multinomialmente sobre tópicos latentes. Además, cada usuario se representa como una distribución de probabilidad sobre los tópicos y cada tópico es una distribución de probabilidad sobre los ítems, que de acuerdo a (Liu, Chen, Member, Xiong, Ding y Chen, 2012), éstos pueden tener múltiples características, que pertenecen a muchos intereses o preferencias latentes. Al mismo tiempo, con los tópicos latentes descubiertos, es posible derivar ítems similares con mayor precisión para comprender las necesidades de los usuarios y hacer recomendaciones más relevantes para ellos (Chang y Hsiao, 2013). La extracción de los intereses del usuario con LDA es un proceso de inferencia de la variable latente, luego de este proceso el valor de estas variables latentes deben maximizar la distribución posterior de la totalidad de los registros de rating del usuario (Liu, Chen, Member, Xiong, Ding y Chen, 2012).

Adicionalmente, con el objeto de clarificar mejor lo analizado en el estado del arte, se presenta un esquema que resume algunos de los elementos involucrados en este trabajo.

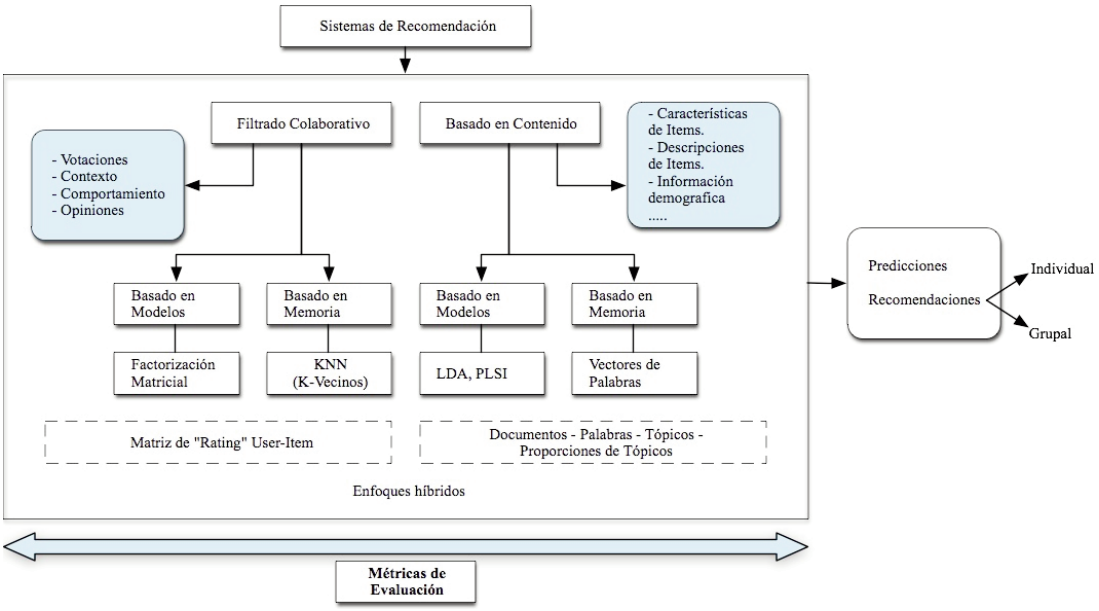

Figura 2. Esquema del estado el arte.

3. Problemas encontrados

De acuerdo a (Adomavicius y Tuzhilin, 2005), el problema principal de la recomendación para sistemas basados en filtrado colaborativo, es la estimación de ratings para los items aún no vistos por el usuario. La mayoría de los sistemas de recomendación se encuentran con el problema de escasez de datos y el problema de arranque en frío (cold-start) (Liu, Wu y Liu, 2013), esto es, cómo proporcionar recomendación a los nuevos usuarios que han expresado muy pocas calificaciones.

Basados en estas premisas, se detallan los problemas identificados en la revisión de la literatura agrupados en base a los temas principals cubiertos en este estudio: enfoques de recomendación y métodos analizados para la recomendación.

A. Enfoques basados en contenido

Con respecto al enfoque basado en contenido se tienen las siguientes limitaciones:

- Análisis de contenido limitado, está relacionado con la eficacia de las palabras claves y con las características asociadas a los ítems, por tanto, las técnicas basadas en el contenido están limitados a las características que se asocian con los ítems que estos sistemas recomiendan (Adomavicius y Tuzhilin, 2005).
- Sobre especialización, dado que los sistemas basados en contenido sólo recomiendan ítems que tienen un alto grado de similitud con aquellos preferidos en el pasado, el usuario está limitado a que el sistema le recomiende ítems que son similares a esos que ya fueron puntuados. Por tanto, el conjunto de ítems recomendados podría ser obvio y demasiado homogéneo (Lops, De Gemmis y Semeraro, 2011).
- Problema de nuevos usuarios, conocido como problema de arranque en frío (cold-start) se asocia con todos los tipos de sistemas de recomendación, pero este tipo de sistemas el tema es particularmente evidente ya que su modelo se basa sólo en valoraciones de los usuarios, es decir que, el usuario tiene que evaluar un número suficiente de ítems antes de que un sistema de recomendación basado en contenido puede realmente entender las preferencias del usuario y presentarle recomendaciones confiables. Esto además significa que los sistemas de recomendación deben ser lo suficientemente capaces para brindar recomendaciones no triviales para un usuario sin suficientes recomendaciones previas en su perfil (Adomavicius y Tuzhilin, 2005).

En este tipo de sistemas los problema de cold-start y de sobre especialzación pueden ser resueltos de manera parcial considerando relaciones sémáticas basados en metadatos de vocabularios semánticos de conceptos (Wang, Stash, Aroyo, Hollink y Schreiber, 2009).

B. Métodos basados en filtrado colaborativo

De acuerdo a (Ahamed y Parambath, 2013; Aghdam, Analoui y Kabiri, 2015) la mayoría de los enfoques actuales de filtrado colaborativo se enfrentan a tres problemas: escasez, escalabilidad y cold-start.

- La escasez de la matriz de puntuación de usuarios- ítems, plantea un problema de modelado en los sistemas de recomendación ya que esto puede dar lugar a graves excesos de ajuste de los datos y resultados con muy mala precisión en la predicción. Este problema de escasez tiene un fuerte efecto sobre el poder predictivo de los algoritmos, y puede conllevar a un sobre ajuste de los datos y dar como resultado una muy mala precisión en la predicción.
- La escalabilidad, en los sistemas recomendadores el número de usuarios e ítems puede ser bastante grande, lo cual puede retardar el proceso de recomendación de manera significativa, principalmente en los métodos basados en memoria los cuales necesitarán bastante computación en este caso. Este problema se deriva del problema de la sobrecarga de información (overload), el cual de alguna manera ya ha sido abordado desde una amplia gama de campos de investigación.
- Arranque en frío (cold-star), cuando la matriz de calificación es escasa, dos usuarios o ítems son poco probable que tenga calificaciones comunes y, en consecuencia, los métodos de recomendación lo que harán es predecir calificaciones utilizando un número muy limitado de vecinos. En vista de que los sistemas de colaboración realizan la predicción basados en la calificación de un usuario similar hacia un ítem, surge el problema de nuevos ítems donde, si un ítem no ha sido calificado por suficientes usuarios, los resultados del sistema de colaboración pueden ser muy sesgados, es decir, no pueden recomendar los ítems que no tienen calificaciones.

En este sentido, los problemas mencionados y los relacionados al uso de los datos que hacen frente a la tarea de recomendación, se han abordado desde una amplia gama de perspectivas que aplican diferentes métodos en orden a proporcionar recomendaciones. Podemos citar entre otros las matrices de factorización, las cuales se han convertido en una de las técnicas principales para dar solución a los problemas de escasez de datos. Sin embargo, debido a la complejidad de tiempo en la composición de recomendaciones, los enfoques basados en estas matrices son ineficientes para hacer frente a una enorme cantidad de datos históricos (Lee y Kwon, 2015). No obstante, los métodos de factorización probabilística pueden ser aplicados a datos masivos y presentan una muy buena escalabilidad. Estos métodos también presentan algunos problemas, por ello a continuación se citan las limitaciones encontradas principalmente con los métodos de recomendación basados en el modelado de factores latentes que son de interés en este estudio.

C. Problemas con Matrices de Factorización

Los métodos de matriz de factorización se consideran de alta escalabilidad, debido a que un método de matriz de factorización de bajo rango supone que existe un pequeño número de factores latentes (características) que pueden explicar el comportamiento del rating de los usuario (Aghdam, Analoui y Kabiri, 2015). A pesar de esto hay dos desventajas principales de las recomendaciones basadas en matrices de factorización (Wang y Blei, 2011):

- El espacio latente aprendido no es fácil de interpretar.
- La factorización de la matriz sólo utiliza información de otros usuarios y no se puede generalizar a ítems completamente no calificados.

En (Gopalan, Ruiz, Ranganath y Blei, 2014), se menciona que una limitación de las matrices de factorización es la selección del modelo. Esto es, escoger el número de componentes con cual modelar los

datos o predecir el desempeño sobre un conjunto de ratings.

En relación a las Matrices de Factorización Probabilísticas, éstas son muy útiles para modelar las clasificaciones de ítems, conocer cuando ocurrió el rating, donde el usuario vive, o que actores aparecen en el item (Adams, Dahl y Murray, 2010). Sin embargo, estos últimos son difíciles de incorporar en el modelo de PMF. Por tanto, una dificultad con el modelo PMF es que a menudo hay más datos disponibles de las observaciones que son necesarias considerar en el modelado.

D. Problemas con el método LDA

Algunos de los problemas identificados entorno al modelo de tópicos en general están relacionados con la representación de tópicos complicados, el no poder capturar variaciones del vocabularios (palabras relacionadas), y la división de palabras ambiguas (Zhao, Cheng, Hong y Chi, 2015). No obstante, los modelos de tópicos probabilístos cubren estas carencias, considerando a los tópicos como una distribiución de palabras, donde multiples palabras permiten describir un tópico complicado, y el peso de las palabras permite modelar variaciones semánticas de un tópico.

Algunos problemas de LDA están relacionados con asumir que tanto el orden de las palabras en el documento y el orden de los documentos en la colección no importan, este segundo supuesto puede ser poco realista en el análisis de las colecciones de larga ejecución. En tales colecciones, podemos querer asumir que los temas cambian con el tiempo. Un enfoque a este problema es el modelado de topicos dinámico, el cual respeta el orden de los documentos y da una estructura de tópicos posterior más rica que LDA (Blei, 2011). Otro problema está relacionado con el número de tópicos fijo y arreglado que se asume en LDA. Algunas soluciones a esto es determinar el número de tópicos durante la inferencia posterior; modelos jerárquicos de tópicos, ente otros.

Otro problema es como mostrar los tópicos y como mostrar mejor un documento con un modelo de tópicos. A nivel de documentos, los modelos de tópicos proporcionan información potencialmente útil sobre la estructura del documento, la cual podría ayudar a los lectores a identificar las partes más interesantes del documento.

4. Cuestiones abiertas identificadas

A parte de los trabajos presentados en el estado del arte, se analizaron algunas revisiones sistemáticas de la literatura entorno a sistemas recomendadores como el presentado por (Park, Kim, Choi y Kim, 2012) donde se mencionan los campos de aplicación de los Sistemas recomendadores y las técnicas de minería de datos utilizadas; en (Bobadilla, Ortega, Hernando y Gutiérrez, 2013) se da una visión general de los sistemas de recomendación, en el que mencionan los métodos de filtrado de contenido, algoritmos utilizados con estos métodos; su evolución, una clasificación original para estos sistemas, áreas de aplicación en el futuro y su importancia. Un estudio sobre desarrollos de sistemas recomendadores, impactos y futuras direcciones en este campos es presentado en (Lü, Medo, Yeung, Zhang, Zhang y Zhou, 2012), los cuales además evalúan algoritmos de filtrado colaborativo, técnicas de reducción de la dimensionalidad para futuros desarrollos. Por otro lado, un estudio sobre el rol de la Matrices de factorización en filtrado colaborativo es el presentado por (Kumar Bokde, Girase y Mukhopadhyay, 2014) donde se presentan un estudio exhaustivo de modelo de matriz de factorización como SVD para abordar los retos de algoritmos CF, que pueden servir como hoja de ruta para la investigación en esta área.

Basados en los estudios más recientes podemos resaltar que:

- El uso de los sistemas de recomendación usando análisis de redes sociales siguen siendo deficientes. Por lo tanto, el desarrollo de la investigación de sistemas de recomendación utilizando análisis de redes sociales será además una interesante área de investigación.

- Algunas futuras direcciones de investigación entorno a LDA están relacionados con: 1) la evaluación del modelo, proponer métodos de evaluación que responden a cómo los algoritmos son utilizados, o cuán interpretables ellos son. 2) Visualización e interfaces: desarrollar nuevos métodos de interacción y visualización de tópicos y corpus (Blei, 2011).
- Otras de las cuestiones abiertas son: la necesidad de ofrecer recomendaciones a grupos de usuarios en lugar de usuarios individuales, y la de abordar las preferencias de los miembros individuales de un grupo de usuarios a fin de proporcionar sugerencias para grupos en su conjunto. Otra importante área identificada es la de proporcionar explicaciones sobre las recomendaciones.

5. Metodología

En este trabajo se propone el uso de un modelo de recomendación que combine modelos probabilísticos gaussianos y de modelado de tópicos para modelar características de filtrado colaborativo y basado en contenido con el objeto de dar solución a algunos de los problemas encontrados tradicionalmente en los sistemas recomendadores.

Se inicia con una revisión comprensiva de los principales aspectos relacionados con sistemas recomendadores para identificar algunas líneas abiertas entorno a ellos. Basado en esto se consideran principalmente las siguientes fases:

1. Identificación de requerimientos para el sistema recomendador
2. Selección de técnicas y métodos que pueden ser adaptadas a un nuevo modelo de recomendación.
3. Pruebas y Evaluación: se prueba el modelo en contextos reales, a fin de medir el rendimiento del mismo. La evaluación del modelo se lleva a cabo utilizando métricas que valoren la calidad de las recomendaciones.

Estas tres fases se integran en un enfoque de desarrollo híbrido que involucra algunas tareas para conseguir los objetivos del trabajo.

6. Análisis del funcionamiento de los modelos de recomendación a utilizar

Para conocer el funcionamiento de los métodos de factores latentes y modelado de tópicos, se realizaron algunas pruebas con la matriz de factorización básica y el método LDA comúnmente utilizado en enfoques basado en contenido, pero que en nuestro caso se lo ajustó para que funcione en un enfoque basado en filtrado colaborativo, considerando para ello la analogía presentada en la figura 1. Para estos experimentos se utilizó el conjunto de datos MovieLens[1], que corresponde a datos de calificaciones de usuarios a películas.

Método	No. Recomendaciones	α	β	Precision (%)	Recall (%)
LDA	10	50/k	0.01	80	13
LDA	10	0.02	0.02	79	14
MF	Diferente para cada usuario	0.0002	0.002	76	77

Tabla 1. Resultados de experimentos.

Con los experimentos realizados con LDA pudimos observar que al variar los valores de los hiperparámteros ($\propto$ y β) el porcentaje de precisión y recall variaba. También se pudo observar que cuando el número de recomendaciones era pequeño se obtenía un valor de precision mayor (Ver tabla 1). Para las experimentaciones se confirguró el número de factores latentes k=20.

Con el método de factorización matricial se encontró el problema que durante el proceso de recomendación se podía dejar sin recomendar un ítem con una puntuación alta. Por lo que, una estrategia que se llevo acabo fue considerar en la recomendación ítems presentes en el subconjunto de evaluación fijando un

[1] MovieLens: http://movielens.org

número de items a recomendar para cada usuario, configurado en base a puntuaciones mayores a un cierto umbral.

Una vez conocido el funcionamiento de estos métodos, se procede a analizar algunos modelos de recomendación que integran estos métodos y que serán utilizados dentro de este trabajo, ya sea como parte de la experimentación en contextos reales o para contrastar con los resultados obtenidos con el modelo híbrido de recomendación probabilístico que se va a ser utilizado en esa investigación, el cual se basa en el trabajo presentado por (Hernando, Jesus y Fernando, 2016), los cuales consideran un modelo de factorización no negativo para sistemas recomendadores basados en modelos probabilísticos bayesianos, utilizado para predecir los gustos del usuatio en sistemas recomendadores de filtrado colaborativo. En este modelo los factores latentess se intrepretan como grupos de usuarios quienes comparten los mismos gustos en el sistema. Por otra parte, los resultados que se obtengan del modelo híbrido serían comparados con los resultados obtenidos de la experimentación del modelo CTR (Collaborative Topic Regression, por sus siglas en inglés), el cual utiliza un algoritmo de recomendación de artículos científicos propuesto por (Wang y Blei, 2011) que combina técnicas de filtrado colaborativo y técnicas y de modelado de topicos. El trabajo presentado por estos autores se basa en el conjunto de datos del sitio CiteUlike[2], donde cada usuario tiene una biblioteca de artículo y la idea es recomendadr artículos de interés que no están en su biblioteca.

7. Conclusiones

Los resultados de la revisión de la literatura pueden proporcionar una orientación útil para profesionales e investigadores en el área. En base a esta revisión se presentan algunos problemas relacionados con los enfoques de recomendación y los métodos mayormente utilizados para el proceso de recomendación. Temas que se consideran relevantes por los desafíos que plantean a la hora de desarrollar sistemas recomendadores y por las innovaciones que suponen tendrían en los contextos que se utilicen.

Dependiendo del enfoque de recomendación es necesario el uso de algún método que se ajuste a éste y que permita seguir el comportamiento del usuario. En este trabajo nos centramos en analizar aquellos métodos que nos permitan alcanzar el objetivo general de la investigación.

Con el objeto de conocer el funcionamiento de las técnicas probabilísticas de fatores latentes y modelado de tópicos, se probaron algunos métodos de recomendación entre ellos matrices de factorización, y LDA, enfocado a sistemas de recomendación de filtrado colaborativo, en el dominio de la recomendación de películas; CTR para recomendaciones basados en contenido y un métodos de recomendación basado en matriz de factorización no negativa.

Agradecimientos

Agradecemos la colaboración del grupo de investigación de Sistemas inteligentes de la UPM y a la sección de inteligencia artificial de la UTPL por el soporte al proyecto de investigación.

Cómo citar este artículo / How to cite this paper

Valdiviezo-Díaz, P.; Hernando, A. (2016). Una comprensiva revisión de los métodos de recomendación basados en técnicas probabilísticas. *International Journal of Information Systems and Software Engineering for Big Companies (IJISEBC)*, 3(2), 65-74. (www.ijisebc.com)

2 CiteUlike: http://www.citeulike.org/

Referencias

Adams, R. P.; Dahl, G. E.; Murray, I. (2010). Incorporating Side Information in Probabilistic Matrix Factorization with Gaussian Processes.
Adomavicius, G.; Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng., 17(6), 734-749.
Aghdam, M. H.; Analoui, M.; Kabiri, P. (2015). A Novel Non-Negative Matrix Factorization Method for Recommender Systems.
Ahamed, S.; Parambath, P. (2013). Matrix Factorization Methods for Recommender Systems.
Benítez, R.; Escudero, G.; Kanaan, S.; Masip, D. (2014). Inteligencia artificial avanzada, Primera Ed. UOC, Barcelona.
Bing, L. (2011). Web Data Mining Exploring Hyperlinks, Contents, and Usage Data, Second Edi. Hardcover.
Blei, D. (2011). Probabilistic topic models. Proc. 17th ACM SIGKDD Int. Conf. Tutorials - KDD '11 Tutorials.
Blei, D.; Ng, A.; Jordan, M. (2003). Latent Dirichlet Allocation.
Bobadilla, J.; Ortega, F.; Hernando, A.; Gutiérrez, A. (2013). Recommender systems survey. Knowledge-Based Syst., 46, 109-132.
Chang, T.; Hsiao, W. (2013). LDA-BASED PERSONALIZED DOCUMENT. In PACIS 2013.
Ekstrand, M.; Riedl, J.; Konstan, J. (2010). Collaborative Filtering Recommender Systems. Found. Trends® Human–Computer Interact., 4(2), 81-173.
Ghauth, K. I.; Abdullah, N. A. (2010). Learning materials recommendation using good learners' ratings and content-based filtering. Educ. Technol. Res. Dev., 58(6), 711-727.
Gopalan, P.; Ruiz, F.; Ranganath, R.; Blei, D. M. (2014). Bayesian Nonparametric Poisson Factorization for Recommendation Systems.
Hernando, A.; Jesus, B.; Fernando, O. (2016). A non negative matrix factorization for Collaborative Filtering Recommender Systems based on a Bayesian probabilistic model. Knowledge-Based Syst., 97, 188-202.
Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. In Proceeding of Uncertainty in Artificial Intelligence, UAI'99.
Huang, Z.; Chung, W.; Hsinchun, C. (2004). A Graph Model for E-Commerce Recommender Systems. J. Am. Soc. Inf. Sci. Technol., pp. 259–274.
Jin Rong, S. L.; Cheng Xiang, Z. (2003). Preference-based Graphic Models for Collaborative Filtering. In Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (UAI 2003), pp. 329-336.
Koren, Y.; Bell, R.; Volinsky, C. (2009). Matrix factorization techniques for recommender Systems. IEEE Comput. Soc., pp. 42-49.
Kumar Bokde, D.; Girase, S.; Mukhopadhyay, D. (2014). Role of Matrix Factorization Model in Collaborative Filtering Algorithm : A Survey Matrix Factorization Model in Collaborative Filtering Algorithms.
Lee, H.; Kwon, J. (2015). Improvement of Matrix Factorization-based Recommender Systems Using Similar User Index.
Liu, J.; Wu, C.; Liu, W. (2013). Bayesian Probabilistic Matrix Factorization with Social Relations and Item Contents for recommendation. Decis. Support Syst., 55(3), 838-850.
Liu, Q.; Chen, E.; Member, S.; Xiong, H.; Ding, C. H. Q.; Chen, J. (2012). Enhancing Collaborative Filtering by User Interest Expansion via Personalized Ranking.
Lops, P.; De Gemmis, M.; Semeraro, G. (2011). Recommender Systems Handbook. Springer US, Boston, MA.
Lü, L.; Medo, M.; Yeung, C. H.; Zhang, Y. C.; Zhang, Z. K.; Zhou, T. (2012). Recommender Systems. Phys. Rep., 519(1), 1-49.
Park, D. H.; Kim, H. K.; Choi, I. Y.; Kim, J. K. (2012). A literature review and classification of recommender systems research. Expert Syst. Appl., 39(11), 10059–10072.
Ricci, F.; Rokach, L.; Shapira, B.; Kantor, P. B. (2011). Recommender Systems Handbook. Springer US, Boston, MA.
Richert, W.; Coelho, L. P. (2013). Building Machine Learning Systems with Python, First. Packt Publishing Ltd, Birmingham.
Salakhutdinov, R.; Mnih, A. (2008). Probabilistic Matrix Factorization. In Neural Information Processing Systems 21 (NIPS 2008).
Wang, C.; Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '11, p. 448.
Wang, Y.; Stash, N.; Aroyo, L.; Hollink, L.; Schreiber, A. T. (2009). Using Semantic Relations for Content-based Recommender Systems in Cultural Heritage. In Proceedings Workshop on Ontology Problems (WOP2009; in conjunction with ISWC2009), pp. 16-28.
Wen, Z. (2008). Recommendation System Based on Collaborative Filtering.
Yu, K.; Zhang, B.; Zhu, H. (2012). Towards Personalized Context-Aware Recommendation by Mining Context Logs through Topic Models.
Zhao, Z.; Cheng, Z.; Hong, L.; Chi, E. H. (2015). Improving User Topic Interest Profiles by Behavior Factorization. In Proceedings of the 24th International Conference on World Wide Web, pp. 1406–1416.