Appendix C

Conclusiones y trabajo futuro

En esta tesis hemos investigado cómo medir y predecir la eficacia de sistemas de recomendación. Hemos analizado y propuesto un conjunto de métodos basados en la adaptación de predictores de eficacia desde el área de Recuperación de Información – principalmente el predictor de claridad de consulta, que captura la ambigüedad de una consulta con respecto a una colección de documentos dada. Hemos definido varios modelos de lenguaje utilizando distintos espacios de probabilidad para capturar los aspectos de los usuarios e ítems implicados en las tareas de recomendación. En este contexto, hemos propuesto y evaluado técnicas novedosas para distintos espacios de entrada extraídas de la Teoría de la Información y la Teoría de Grafos Sociales, usando propiedades sobre las preferencias de usuario así como métricas de grafos, como PageRank sobre la red social del usuario.

Más aún, dado que queremos predecir la eficacia de un sistema de recomendación particular, necesitamos una metodología de evaluación clara con la cual las predicciones de eficacia puedan ser contrastadas. Así, en esta tesis investigamos la metodología de evaluación como parte del problema abordado, donde hemos identificado sesgos estadísticos en la evaluación de la recomendación – a saber, sesgos de dispersión en test y popularidad – los cuales pueden distorsionar las medidas de eficacia, y por tanto, confundir el poder aparente de los métodos de predicción de eficacia. Hemos analizado en profundidad el efecto de dichos sesgos, y hemos propuesto dos diseños experimentales capaces de neutralizar el sesgo de popularidad: una técnica basada en percentiles y un test uniforme. El análisis sistemático de las metodologías de evaluación y las nuevas variantes propuestas permiten una valoración más completa y precisa de la eficiencia de nuestros métodos de predicción de eficacia.

Por otro lado, hemos explotado los métodos propuestos de predicción de eficacia en dos aplicaciones donde se usan para ponderar dinámicamente distintos componentes de un sistema de recomendación, a saber, el ajuste dinámico de recomendaciones híbridas ponderadas, y la ponderación dinámica de las preferencias de los vecinos en filtrado colaborativo basado en usuario. A través de una serie de experimentos empíricos con varios conjuntos de datos y diseños experimentales, hemos encontrado una correspondencia entre el poder predictivo de nuestros predictores de eficacia y la mejora en eficacia de las dos aplicaciones evaluadas.

Presentamos aquí las conclusiones principales obtenidas en este trabajo de investigación. La Sección C.1 muestra un resumen y discusión de nuestras contribuciones, mientras que en la Sección C.2 mostramos vías de investigación que puedan ser abordadas como trabajo futuro.

C.1 Resumen y discusión de las contribuciones

En las siguientes subsecciones resumimos y discutimos las principales contribuciones de esta tesis, abordando los objetivos enunciados en el Capítulo 1. Estas contribuciones están organizadas de acuerdo a los tres objetivos principales de la tesis. Primero hemos analizado cómo evaluar adecuadamente los sistemas de recomendación para obtener medidas no sesgadas de su eficacia. Segundo hemos propuesto predictores de eficacia que tratan de estimar la eficacia de un método de recomendación. Y tercero hemos usado los predictores de eficacia propuestos para combinar dinámicamente componentes de un sistema de recomendación.

C.1.1 Análisis de la definición y evaluación de la eficacia en sistemas de recomendación

Hemos analizado distintas alternativas de diseño experimental disponibles en la literatura para sistemas de recomendación, orientados, en particular, a la evaluación basada en rankings, y hemos mostrado que **las hipótesis y condiciones subyacentes al paradigma Cranfield no se pueden asumir en los entornos habituales de recomendación**. Específicamente, hemos detectado sesgos estadísticos que aparecen al aplicar dicho paradigma a la evaluación de sistemas de recomendación. Hemos mostrado que el valor específico de la métrica de evaluación es útil en términos comparativos, pero no tiene un sentido particular en términos absolutos. Hemos mostrado que la precisión decrece linealmente con la dispersión de los ítems relevantes (**sesgo de dispersión**) al usar la metodología de evaluación AR, mientras que no sufre de este sesgo al usar la estrategia 1R.

También hemos observado que un algoritmo de recomendación no personalizado basado en la popularidad de los ítems obtiene valores de eficacia altos, y hemos mostrado y analizado en detalle cómo y por qué esto es debido a un **sesgo de popularidad** en la metodología experimental. Para abordar estos problemas, en esta tesis hemos propuesto **técnicas experimentales novedosas que neutralizan satisfactoriamente el sesgo de popularidad**.

C.1.2 Definiciones y adaptaciones de predictores de eficacia para sistemas de recomendación

En esta tesis hemos definido y elaborado **predictores de eficacia en el contexto de recomendación**, normalmente tomando al usuario como el objeto de la predicción, pero también considerando los ítems como entradas alternativas para la predicción. Específicamente, hemos adaptado el predictor de eficacia de consulta conocido como *claridad* tomando distintas hipótesis y formulaciones para obtener diferentes variaciones

de los predictores de **claridad del usuario**. También hemos usado conceptos relacionados con la Teoría de la Información como la entropía, métricas de grafos como la centralidad, PageRank y HITS, y otras técnicas heurísticas y específicas del dominio. Hemos definido estos predictores basándonos en tres espacios de entrada para las preferencias de los usuarios: **puntuaciones, registros** y **redes sociales**. Sobre puntuaciones y registros hemos definido varios modelos de lenguaje y espacios de vocabulario de tal manera que nuestras adaptaciones de claridad capturen distintos aspectos del usuario en una formulación unificada para ambos espacios de entrada. Dentro del mismo marco, hemos introducido la dimensión temporal en los datos de preferencia basados en registros, considerando y elaborando predictores de eficacia basados en tiempo propuestos en trabajos sobre búsqueda ad-hoc previos en el área de RI.

Además, hemos definido **predictores basados en ítems** cuando se usan preferencias basadas en puntuaciones, los cuales intentan estimar la eficacia de los objetos en consideración (siendo más precisos, la eficacia de un sistema de recomendación cuando sugiere dichos ítems). Aquí el principal problema consiste en cómo definir una métrica de eficacia real de manera que un predictor intente estimarla, ya que los ítems no son la entrada principal del proceso de recomendación. Por esta razón, hemos desarrollado metodologías novedosas donde la eficacia de un ítem pueda ser medida, también considerando posibles sesgos que pudieran aparecer cuando usuarios con muchas puntuaciones pueden distorsionar los resultados por razones estadísticas.

Hemos evaluado el acierto predictivo de nuestros métodos calculando la correlación entre la eficacia estimada y la real, siguiendo la práctica estándar en la literatura de predicción de eficacia en RI. De esta manera, hemos usado las metodologías no sesgadas analizadas a lo largo de esta tesis para **comparar cómo se comportan los predictores cuando los sesgos de dispersión y popularidad han sido neutralizados**. Hemos encontrado fuertes valores de correlación confirmando que nuestras técnicas muestran un **poder predictivo significativo**.

C.1.3 Ponderación dinámica en sistemas de recomendación

La combinación de algoritmos de recomendación es frecuente en la literatura de los Sistemas de Recomendación, en especial lo que se conoce como conjuntos de algoritmos de recomendación (*ensembles*), que son un tipo particular de métodos de recomendación híbrida donde se combinan varios algoritmos, y que actualmente son muy comunes en el área, tal y como se puede comprobar en competiciones recientes (Bennett and Lanning, 2007; Dror et al., 2012). El filtrado colaborativo, una de las técnicas más usadas dentro de la colección de estrategias de recomendación, también se puede ver como una combinación de varias subfunciones de utilidad, cada una correspondiendo a un vecino (en un filtrado basado en usuario). De la misma manera

que los predictores de eficacia en Recuperación de Información se han usado para optimizar la agregación de rankings, nosotros hemos investigado el uso de predictores de eficacia en recomendación para agregar dinámicamente la salida de los algoritmos de recomendación y los vecinos.

Hemos definido un **marco de hibridación dinámica** donde los conjuntos de algoritmos de recomendación pueden beneficiarse de las ponderaciones dinámicas de acuerdo a los predictores de eficacia con los que muestran correlaciones altas. Nuestros resultados indican que correlaciones altas con la eficacia tienden a corresponder con mejoras en los algoritmos de recomendación híbridos dinámicos. Además, los conjuntos dinámicos de recomendación han mostrado mejor eficacia que los conjuntos estáticos para distintas combinaciones de algoritmos y en los tres tipos de predictores de eficacia investigados.

Por otro lado, también hemos propuesto un marco para la **selección y ponderación de vecinos** en sistemas de filtrado colaborative basados en usuario. Hemos definido predictores y métricas de eficacia de vecinos adaptando e integrando algunos de los métodos de la literatura en recomendación sensibles a la confianza de usuarios. Nuestro marco unifica varias nociones de eficacia de vecinos bajo la misma forma, y presenta un análisis objetivo del poder predictivo de diferentes funciones de valoración de vecinos. Una vez el poder predictivo de estos predictores de vecinos ha sido confirmado, usamos dichos métodos para ponderar la información que proviene de cada vecino de manera dinámica, experimentando con distintas estrategias para la combinación de valores de similitud y pesos de los vecinos. Nuestros experimentos confirman una correspondencia entre los análisis de correlación y los resultados finales de eficacia, en el sentido de que los valores de correlación obtenidos entre los predictores y las métricas de eficacia de vecinos anticipan qué predictores obtendrán mejor eficacia cuando se introduzcan en un algoritmo de filtrado colaborativo basado en usuarios.

C.2 Trabajo futuro

La predicción de eficacia en recomendación es un área interesante de investigación también desde una perspectiva de negocio, ya que podríamos decidir cuándo entregar las recomendaciones al usuario, evitando disminuir la confianza de los usuarios sobre la relevancia de las sugerencias del sistema. En este sentido, las predicciones pueden dar un control al proveedor de servicios, un control que podría usarse potencialmente de varias maneras, incluyendo una combinación de métodos más general que la que se ha abordado en esta tesis. Independientemente de cualquier aplicación plausible para la industria, y más allá de los logros presentados a lo largo de esta tesis, contemplamos las líneas de investigación futuras que describimos a continuación.

La evaluación de los sistemas de recomendación es aún un objeto de investigación activa en el campo, donde varias cuestiones requieren atención, como el vacío entre experimentos en línea (*online*) y de fuera de línea (*offline*). No obstante, en esta tesis hemos enfocado nuestra investigación en aspectos relacionados con la predicción de eficacia, lo cual requiere un conocimiento más profundo de las metodologías de evaluación utilizadas. De esta manera, podríamos **extender nuestro análisis de las metodologías de evaluación a otras métricas de ranking**, como a aquellas basadas en dos listas de recomendaciones (NDPM y las correlaciones de Spearman y Kendall) o a aquellas adaptadas de Aprendizaje Automático (por ejemplo, el área bajo la curva o AUC en inglés). De esta manera, podríamos encontrar que alguna de estas métricas no está influida por ninguno de los sesgos descritos en el Capítulo 4, o que ninguno de los diseños alternativos propuestos es capaz de neutralizar esos efectos. Como un ejemplo del interés de este tema, recientemente en (Pradel et al., 2012) los autores analizaron los efectos de popularidad sobre la métrica AUC, y encontraron que considerar los datos no puntuados como información negativa durante el entrenamiento podría mejorar la eficacia, pero también podría favorecer a los algoritmos de recomendación basados en popularidad con respecto a los personalizados.

Además, sería beneficioso para nuestra investigación ser capaces de validar la utilidad de las medidas no sesgadas de eficacia con evaluaciones en línea. Esto sería valioso para tener una valoración comparativa con las observaciones fuera de línea que hemos obtenido, así como un conocimiento más profundo de la magnitud por la cual la popularidad puede ser o no una señal ruidosa. Tal estudio de usuario nos ayudaría a determinar los beneficios reales (si los hubiera) de recibir recomendaciones populares, ya que, por ejempo, por definición estas sugerencias no serían novedosas ni probablemente causales o diversas.

En el Capítulo 6 hemos propuesto varios predictores de eficacia para recomendación basados en los mismos principios de aquellos denominados como predictores pre-búsqueda en Recuperación de Información, como la claridad, donde la salida del algoritmo de búsqueda (o de recomendación en nuestro caso) no es usada por el predictor. Teniendo en cuenta nuestros resultados, las posibilidades para investigar más predictores de eficacia en recomendación son abundantes. En esta línea, varios autores han explotado la **combinación de predictores para obtener valores de correlación mayores y un poder predictivo mayor**, como (Hauff et al., 2009) y (Jones and Diaz, 2007), donde se han usado regresión penalizada y regresión lineal con aprendizaje mediante redes neuronales, respectivamente. En esos trabajos la combinación de predictores de distinta naturaleza mejoró la correlación con respecto a una métrica de evaluación objetivo – en este caso, la precisión promedio. Por ello, creemos que la combinación de predictores puede ser válida también para recomendación, especialmente sabiendo que hemos definido predictores basados en diferentes tipos de datos de los que se espera baja redundancia entre ellos y, por tanto, que dicha

combinación pueda producir correlaciones mayores. Ejemplos de tales combinaciones podrían ser la mezcla de dimensiones sociales y temporales, los predictores temporales basados en ítems, u otras dimensiones contextuales no abordadas en esta tesis.

Más aún, una futura investigación podría **analizar y adaptar también a los sistemas de recomendación predictores de eficacia post-búsqueda** definidos en la literatura de RI, como por ejemplo aquellos basados en el análisis de la distribución de las puntuaciones de los ítems recomendados a cada usuario. Esto podría conseguir predictores con correlaciones más fuertes y, por tanto, con mayor poder predictivo de la eficacia de los algoritmos de recomendación, como ocurre en RI donde los predictores post-búsqueda normalmente obtienen valores de correlación mayores que los pre-búsqueda. La principal limitación de este tipo de predictores es que no pueden ser usados directamente para adaptar la salida de los algoritmos de recomendación, ya que normalmente se requiere la salida completa – es decir, el ranking – para el cálculo de los valores del predictor. Esto obligaría a pensar en distintas aplicaciones donde este tipo de predictores pudieran ser usados en recomendación.

Una dirección particular digna de ser considerada y también relacionada con el Capítulo 6, sería el **uso de técnicas de evaluación alternativas** más allá de las métricas de correlación, como aquellas basadas en el agrupamiento entre los valores de eficacias reales y estimados (ver Sección 5.4.2). En nuestro trabajo nos hemos centrado en el uso de métricas de correlación, principalmente la correlación de Pearson. Estas métricas tienen limitaciones bien conocidas, como su sensibilidad a los valores extremos y correlaciones no significativas cuando se usan un número pequeño de puntos. Por esta razón, se han propuesto otras técnicas para evaluar el poder predictivo de los predictores. Hemos de notar, sin embargo, que el uso de una técnica particular de evaluación debería enfocarse a su aplicación en contextos específicos (Pérez-Iglesias and Araujo, 2010); en particular esto requiere la definición de nuevas aplicaciones para predictores de eficacia que encajen con la métrica de evaluación, lo cual también contemplamos como un potencial trabajo futuro.

También en el Capítulo 6 hemos desarrollado una metodología de evaluación para estimar el valor real de eficacia de los ítems, con el objetivo de evaluar los predictores de ítems propuestos. Esta metodología debería ser validada para **obtener una medida justa de la eficacia del ítem**, lo cual en este momento es aún un problema abierto. De esta manera, seríamos capaces de definir predictores de ítems adicionales para otros espacios de entrada además de las puntuaciones, y de mejorar la capacidad de predicción de los actuales predictores de eficacia de ítems.

En el Capítulo 7 presentamos experimentos sobre combinación dinámica de algoritmos de recomendación en conjunto. Esos experimentos estaban limitados a un único predictor de eficacia por cada par de algoritmos, así que pretendemos extender dichos experimentos con **conjuntos de algoritmos de recomendación donde se consideren dos predictores** para investigar qué condiciones deberían satisfacerse

entre cada par de predictores de manera que se mejore la eficacia del conjunto. Una vía de investigación relacionada a considerar sería el análisis de valores de correlación tales que se obtengan buenos resultados de eficacia en los métodos híbridos dinámicos. Más específicamente, queremos saber si es mejor tener un fuerte valor de correlación en general (en promedio) o un valor medio no tan fuerte pero mejores estimaciones para usuarios particulares, que tendrían un papel significativo en el sistema similar a los usuarios poderosos (*power users*) definidos en (Lathia et al., 2008). En ese punto, se podría realizar un estudio como el presentado en (Hauff et al., 2010), donde simulaciones de predictores con distintos valores de correlación son evaluados, y cuyos efectos sobre la eficacia final en conjuntos de algoritmos de recomendación son comparados.

Además, otra limitación de los experimentos presentados en el Capítulo 7 es que el tamaño de los conjuntos siempre es dos. Pretendemos considerar **conjuntos de algoritmos de recomendación de tamaño N** y a la larga, como se mencionó antes, usar un predictor de eficacia para cada algoritmo de recomendación. Este es un paso natural, pero no trivial hacia la generalización del marco propuesto de conjuntos completos de algoritmos de recomendación. De manera alternativa, podrían usarse técnicas de Aprendizaje Automático para aprender los mejores pesos a usar por cada usuario e ítem en el conjunto. En este caso, se debería investigar un compromiso entre los costes computacionales de cada técnica (aprendizaje automático frente a predictores de eficacia), su poder predictivo y la tendencia a sobreajustar los datos.

Finalmente, en el Capítulo 8 investigamos el problema de ponderación dinámica de vecinos usando predictores de eficacia de vecinos orientados a métricas de error. El trabajo futuro relacionado con este capítulo podría centrarse en la **adaptación de las métricas de eficacia de vecinos usadas en nuestra propuesta hacia métricas de ranking**, tales como la precisión y el *recall*. Como ya hemos discutido, las métricas de error no son la mejor manera de medir la eficacia, aunque se pueden considerar apropiadas en este contexto ya que queremos medir la mejora en la exactitud de nuestras propuestas, y a la vez facilitar comparaciones con el estado del arte en recomendación sensible a la confianza, donde estas métricas son predominantes. Por lo tanto, el uso de métricas de ranking sería una valiosa contribución al campo por sí misma. Además, una vez tuviéramos definidas métricas de eficacia de vecinos basadas en ranking, seríamos capaces de medir la correlación de los predictores de eficacia de vecinos descritos en este capítulo con tales métricas, y analizar en detalle su poder predictivo con métricas de ranking. Idealmente, podríamos obtener un predictor con suficiente poder predictivo usando los dos tipos de métricas de eficacia de vecinos (las basadas en error y en ranking), aunque esto no es fácil de garantizar en general, ya que cada métrica se define para optimizar distintos parámetros y conceptos.