Crítica: Evaluating Recommendation Systems

En este paper se exploran distintas características atractivas para maximizar con sistemas recomendadores. Cada una de estas características son necesarias en sistemas que quieran hacer predicciones. Por ejemplo, una página de libro busca que sus recomendaciones tengan alta precisión, pero además buscan que las recomendaciones no sean monótonas si no que diversas, que estas sorprendan al usuario porque de ninguna otra manera el usuario la podría haber encontrado, y que muestren una gran gama de su. Además, vemos en detalle cómo son las estructuras de distintos tipos de experimentos, en donde se pueden obtener distintos resultados y de distinta calidad, pero a precios y consecuencias distintas.

La primera idea que me surgió durante la lectura fue si existe un patrón con respecto a resultados en experimentos offline o de estudio de usuarios, y los resultados en experimentos de gran escala. Aunque no todas, distintas características pueden ser evaluadas empíricamente a partir de estudios offline o de estudios de usuarios. Sin embargo, me imagino que no todas las características se traducen igual de bien a un modelo real, corriendo a gran escala. Así, me pregunto si habrá distintos criterios para aceptar distintos experimentos dependiendo específicamente en qué características están midiendo.

Además, encontré muy interesante la idea de utilizar los timestamps de ratings para simular comportamientos de usuario. Me gustaría ver cómo se puede complejizar más el modelo, agregando tendencias de comportamiento dado las épocas. Como hemos visto, ciertos ratings dependen del tiempo (por ejemplo, películas de navidad). Así, se podrían buscar patrones que indiquen que una película tendrá buenos ratings dado las fechas en donde sus ratings ocurrieron, y tener una mejor predicción de comportamientos de usuario de forma más general, asemejándose a casos que ocurren durante todo el año.

Dos dudas me quedaron pendientes. Primero, se dice que la precisión (accuracy) de un modelo es independiente a la interfaz de usuario, por lo que se puede estudiar en un experimento offline. Sin embargo, existen casos donde puedo dar varias recomendaciones y me basta con que el usuario elija una para que el sistema funcione. Por ejemplo, Netflix puede entregar 5 recomdaciones para “Watch Next”, pero solo le importas que elijas una, a diferencia de Amazon que puede recomendar 5 productos para venderte, y le interesa vender los 5. Así, importa cuántos ítems soy capaz de mostrar en mi interfaz sin que cada ítem se muestre de manera muy poco descriptiva. La segunda tiene que ver con la justificación para asignar un peso de 0.5 a la predicción empatadas en la ecuación 12. Lo encontré como un número arbitrario y que posiblemente se pueda encontrar su valor óptimo de manera empírica.

Por último, no estoy de acuerdo con una de las reclamaciones asumidas. En un momento se menciona que, si una usuaria elige los ítems 1,3 y 10, significa que a ella no le interesaron los

ítems 2,4,5,6,7,8 y 9, y que, si además sabemos que abrió la siguiente pestaña mostrando las siguientes 10 recomendaciones, y no eligió ninguna, tampoco encontró interesante esas recomendaciones. Sin embargo, yo pensaría la probabilidad de abrir una recomendación dado que ya abrí varias antes disminuye drásticamente con el número de recomendaciones abiertas con anterioridad. Así, la selección (o su ausencia) de un ítem me dice sobre mi usuario en cierta proporción a su posición dentro de mis recomendaciones.

Alejandro Quiñones