PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
Departamento de Ciencias de la Computación
IIC3633 - Sistemas Recomendadores
Segundo Semestre 2018

Review 3
Evaluating Recommendation Systems

Genaro Laymuns

El paper trata sobre todos los pasos a seguir para evaluar y comparar distintos sistemas recomendadores. Este proceso consiste fundamentalmente de tres etapas: en primer lugar se utilizan experimentos *offline* (fáciles de manejar, pues no requieren interacción con el usuario); luego se realizan experimentos con un grupo específico de usuarios para que entreguen información en cuanto a su experiencia; finalmente se pone en uso el sistema recomendador para un grupo más grande de usuarios, los que generalmente no están conscientes del experimento (más difícil obtener información, pero mucho más cercano a la realidad).

En cuanto a los experimentos *offline*, es importante preguntarse la técnica que se utilizará para definir el *training dataset*. Hay que tener en cuenta que si se eligen muy pocos datos para conformar el *training dataset*, se estará beneficiando a modelos que funcionen mejor con una base de datos *sparse*. Por otro lado, si cada evaluación tiene la hora en que se realizó la evaluación, una buena manera de separar el *dataset* es elegir una hora específica y decir que todas las evaluaciones realizadas antes de ese horario conformarán el *training dataset* y que las demás serán el *testing dataset*. Si bien esto es una buena aproximación, hay que tener en cuenta que los datos fueron generados con un sistema recomendador específico, y la decisión de los usuarios puede verse afectada por las recomendaciones del sistema en funcionamiento. Por otro lado, siempre debemos tener presente la complejidad de nuestro algoritmo y qué tan fácil es testearlo con nueva información, pues si bien un algoritmo puede estar optimizado para funcionar bien en la base de datos, podría ocurrir que su complejidad haga muy difícil (o incluso haga inviable) el testeo con un grupo de usuarios real.

Una vez elegido un sistema recomendador con buenos resultados en la experimentación *offline*, es llevado a la segunda etapa, la cual consiste de experimentación con usuarios conscientes del testeo. Esta etapa siempre estará sesgada, pues no son usuarios reales que consumen un producto, sino que al estar conscientes del experimento pueden alterar su evaluación (por ejemplo, para dejar "más contenta.ª la empresa) y usualmente es muy cara de implementar, pues se necesita un número significativo de usuarios para obtener un buen *feedback* del método.

Al momento de evaluar un sistema recomendador, debe tenerse mucho cuidado al momento de ver su efectividad, puesto que en una clásica tabla de verdad se consideraría como pérdida los productos no recomendades pero que el usuario sí habría utilizado (falsos negativos). Al momento de ver la efectividad de un sistema recomendador en un esquema *offline*, este número está altamente sobredeterminado, pues en la vida real, es muy probable que un usuario no consuma dicho producto si el sistema no se lo recomendó. En el caso de Netflix esto es muy común, ya que usualmente los usuarios escogen una serie/película que se encuentre dentro de las primeras tres categorías, y por lo tanto suelen consumir los productos que les son recomendados.