Sistemas de Recomendación

v. 1.0.0

Uayeb Caballero
uayeb.caballero@gmail.com

Universidad Nacional Autonoma de Honduras

July 17, 2017

Introduction

Los sistemas de recomendaciones son herramientas que generan recomendaciones sobre un determinado objeto de estudio, a partir de las preferencias y opiniones dadas por los usuarios. El uso de estos sistemas se está poniendo cada vez más de moda en Internet debido a que son muy útiles para evaluar y filtrar la gran cantidad de información disponible en la Web con objeto de asistir a los usuarios en sus procesos de búsqueda y recuperación de información. En este trabajo realizaremos una revisión de las características y aspectos fundamentales relacionados con el diseño, implementación y estructura de los sistemas de recomendaciones analizando distintas propuestas que han ido apareciendo en la literatura al respecto.

Collaborative Filtering

El Filtrado colaborativo (FC) es una técnica utilizada por algunos sistemas recomendadores. En general, el filtrado colaborativo es el proceso de filtrado de información o modelos, que usa técnicas que implican la colaboración entre múltiples agentes, fuentes de datos, etc. Las aplicaciones del filtrado colaborativo suelen incluir conjuntos de datos muy grandes. Los métodos de filtrado colaborativo se han aplicado a muchos tipos de datos, incluyendo la detección y control de datos (como en la exploración mineral, sensores ambientales en áreas grandes o sensores múltiples, datos financieros) tales como instituciones de servicios financieros que integran diversas fuentes financieras, o en formato de comercio electrónico y aplicaciones web 2.0 donde el foco está en los datos del usuario, etc. Esta discusión se centra en el filtrado colaborativo para datos de usuario, aunque algunos de los métodos y enfoques pueden aplicarse a otras aplicaciones.

Índice Jaccard

El índice de Jaccard (IJ) o coeficiente de Jaccard (IJ) mide el grado de similitud entre dos conjuntos, sea cual sea el tipo de elementos.
La formulación es la siguiente:
$J(A,B) = |A \cap B| / |A \cup B|$

Basado en memoria

Este mecanismo utiliza los datos de las evaluaciones de los usuarios para calcular la similitud entre los usuarios o elementos. Esto se utiliza para hacer recomendaciones. Este fue de los primeros mecanismos y se usa en muchos sistemas comerciales.

$$\text{sim}(x, y) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| \times ||\vec{y}||} = \frac{\sum_{i \in I_{xy}} r_{x,i} r_{y,i}}{\sqrt{\sum_{i \in I_x} r_{x,i}^2} \sqrt{\sum_{i \in I_y} r_{y,i}^2}}$$

Reducción de Dimensiones

Para conjuntos de datos de alta dimensión (es decir, con número de dimensiones más de 10), reducción de la dimensión se realiza generalmente antes de la aplicación de un K-vecinos más cercanos (k-NN) con el fin de evitar los efectos de la maldición de la dimensionalidad.
Extracción de características y la reducción de la dimensión se puede combinar en un solo paso utilizando análisis de componentes principales (PCA), análisis discriminante lineal (LDA), o análisis de la correlación canónica (CCA) técnicas como un paso pre-procesamiento seguido por la agrupación de K-NN en vectores de características en el espacio reducido dimensión. En aprendizaje automático este proceso de pocas dimensiones también se llama incrustar.

Descomposición en valores singulares

En álgebra lineal, la descomposición en valores singulares de una matriz real o compleja es una factorización de la misma con muchas aplicaciones en estadística y otras disciplinas.
Dada una matriz real $A \in \Re^{m \times n}$, los autovalores de la matriz cuadrada, simétrica y semidefinida positiva $A^T A \in \Re^{n \times n}$ son siempre reales y mayores o iguales a cero. Teniendo en cuenta el producto interno canónico vemos que:

$\left(A^T A\right)^T = A^T \left(A^T\right)^T = A^T A$. O sea que es simétrica
$(Ax, Ax) = x^T A^T Ax = ||Ax||^2 \geq 0$ es decir $A^T A$ es semidefinida positiva, es decir, todos sus autovalores son mayores o iguales a cero.

Si λ_i es el i-ésimo autovalor asociado al i-ésimo autovector, entonces $\lambda_i \in \Re$. Esto es una propiedad de las matrices simétricas.

Conclusión

La cantidad de algoritmos para construir sistemas de recomendacion son incontables, hoy en dia con el constante crecimiento de Internet y con las distintas necesidades para retener los usuarios, los recommendation engines deben de ser mas precisos hasta el punto de dar la experiencia de personalizacion por usuario. tener este tipo de paradigmas hace que el costo computacional sea elevado en el caso de querer implementar tecnicas en base a filtros colaborativos, por lo que para datasets pequeños se recomienda su uso. lo bueno de este tipo de tecnicas es que su interpretabilidad es mas rapida. para el caso de datasets muy grandes se recomiendan tecnicas que involucran reduccion de dimensiones como el caso de Descomposicion Espectral, SVD o ACP. Estos por su tecnica son mas rapidos de procesar pero mas dificiles de interpretar.