Recuperación de Información y Recomendaciones en la Web

Entrega Final

María Jimena Campiotti	4.898.738-9
Guzmán Perera	4.471.271-6
María Belén Taboas	4.628.815-5

Índice

1. Introducción

En vista de que los sitios web del sector de cines tienen grandes carencias al momento de realizar búsquedas y filtrar preferencias de películas, se planteó la necesidad de crear una web amigable al usuario que mejore el flujo de interacción.

En este proyecto, se busca optimizar la búsqueda de películas para el usuario y personalizarla a través de la extracción de información de la web de forma automatizada utilizando técnicas de Web Scraping.

a. Problema a abordar

Hoy en día existen diversos cines, cada uno con sus respectivos complejos. A su vez, cada uno de estos complejos exhibe una gran cantidad de películas. La gran cantidad de oferta existente puede generar problemas a los usuarios a la hora de buscar y definir qué película efectivamente quieren ver y en dónde. Por ejemplo, si se desea saber a qué hora dan cierta película (independientemente del cine) no hay otra forma de saberlo que ingresando a la página web de cada cine recabando dicha información, lo que puede resultar engorroso para cualquier persona.

Por esta razón, la idea que se plantea es que toda esta información se encuentre centralizada en un mismo lugar, de forma que el usuario no deba ingresar a todas las páginas web por separado para encontrar lo que busca. Es decir, homogenizar la información de todos los cines presentados. Además, se busca que la información sobre horarios y películas pueda ser filtrada según las preferencias del consumidor.

b. Solución existente

Existe una solución en la actualidad que cubre parte del problema que se quiere resolver. La página web que lo contempla es www.cartelera.com.uy. La misma cuenta con información sobre los horarios de las películas de todos los cines del país. Sin embargo, la página no cuenta con la opción de aplicar filtros sobre la búsqueda de información. Por ejemplo, en dicha página solamente se puede ordenar la búsqueda de manera alfabética o por calificación. La idea es que la nueva solución propuesta no solo permita ver los horarios de las películas que se encuentran hoy en día en cartelera, sino que también el usuario pueda aplicar filtros de búsqueda según diversos criterios. De esta manera, se priorizaría la usabilidad y la interacción con el usuario, algo que la solución existente no contempla.

2. Desarrollo

a. Análisis de fuentes

A continuación, se pasa a detallar el análisis realizado para lograr recuperar la información provista por las páginas web.

Para poder realizar el trabajo se debió realizar un fuerte análisis de fuentes de distintas páginas de cines. De manera de modelar la información recolectada se debía examinar qué información era proporcionada en cada página y cómo la misma era desplegada, para poder extraerla.

Durante este proceso, se detectó que cada empresa de cine exponía la información de formas distintas. Esto traía problemas al momento de tratar de unificar los datos de las distintas fuentes, ya que tanto el formato como los nombres de los atributos a recuperar eran diferentes. Otro problema que se observó fue la falta de estandarización. Es decir, no se encontraban los mismos campos o características de las películas. Por ejemplo, un complejo mostraba los datos del director, año y país de una película, mientras que otro presentaba solo director y elenco. Esto generaría una dificultad importante al momento de querer normalizar los datos obtenidos.

Los problemas mencionados en el párrafo anterior requerían implementar un scrapper diferente para cada página web de cada cine. Sin embargo, esta solución requeriría un esfuerzo muy grande de desarrollo, además de no ser escalable, ya que, al aumentar la cantidad de cines, se debía agregar un nuevo scrapper por cada uno.

A raíz de los inconvenientes encontrados, se optó por seleccionar la información de una sola página: www.cartelera.com.uy. Dicha web unifica los datos de películas exhibidas en diversos cines del país. Esto presenta varias ventajas: para empezar, contiene la misma información para todas las películas que allí se encuentran. Es decir, los datos mantienen un formato estándar y normalizado. De todas formas, una desventaja de esta solución es que no se obtiene la información directamente de los cines, sino que son datos de segunda mano, por lo que se depende de que esa información sea actualizada para que la información obtenida sea correcta.

En conclusión, se recuperará toda la información proporcionada por la página de "Cartelera" y se aplicarán filtros a los datos obtenidos de manera de proporcionar una mejor experiencia de usuario a los clientes finales.

b. Descripción de componentes en alto nivel

El sistema cuenta con tres módulos que interactúan para poder brindar las funcionalidades necesarias. Éstos son:

i. Back-End

Scraper: encargado de obtener la información necesaria de las páginas web mediante técnicas de Web Scraping. En particular, se buscan datos asociados a las películas junto con sus horarios y el complejo en donde se exhibe cada una.

Persistencia: encargado de persistir la información obtenida por el módulo de Scraping en una base de datos. Este módulo es consumido por el Scraper, anteriormente descripto, para almacenar los datos recabados. Además, este componente, se encarga de realizar las consultas necesarias a la base de datos para poder recuperar la información previamente almacenada. Estas funcionalidades son utilizadas por los servicios web para obtener los datos y luego desplegarlos en el front-end.

ii. Front-End

Interfaz Web: es el componente que provee la interacción entre el usuario y el sistema. Este módulo fue implementado en AngularJS y se encarga de consumir servicios web RESTful para obtener los datos requeridos por el usuario. Además, despliega la información recuperada de manera optimizada y amigable para el usuario.

c. Herramientas

A continuación, se describen las herramientas que se utilizaron para la implementación del sistema desarrollado.

Para el Front-End se utilizó AngularJS[1]. Éste es un framework de JavaScript para la creación y mantenimiento de aplicaciones web.

Por otro lado, para la implementación del Back-End se utilizó el lenguaje de programación Java. El mismo cuenta con una librería denominada jsoup[2], que permite manipular código HTML para realizar el web scraping y obtener la información necesaria.

Finalmente, para la persistencia, se utilizó una base de datos MySQL. El acceso a la base de datos se realizó a través de la API JDBC[3].

d. Análisis de Filtros

La propuesta de valor del proyecto presentado se basa en los filtros que el usuario puede utilizar para realizar sus búsquedas. Es por eso que se hizo énfasis en los mismos y se seleccionaron los filtros que le serán de mayor utilizad al usuario.

[1] https://angularjs.org
[2] https://jsoup.org
[3] https://docs.oracle.com/javase/tutorial/jdbc/

Un primer filtro elegido es por complejo. Este filtro logra que se le muestre al usuario solo aquellas películas que sean exhibidas en el complejo seleccionado. De esta manera, si el usuario desea ir a cierto lugar puede seleccionar de las funciones correspondientes. Junto con este filtro, también se puede realizar un filtrado más general y filtrar por Cine. En este caso cuando se dice Cine se refiere a la organización, como son Movie, Life Cinema, GrupoCine, para nombrar algunos. Por lo que se mostrarían todas aquellas películas que sean exhibidos en todos los complejos de la organización.

Un segundo filtro es por género. Con este filtro se pretende que el usuario obtenga las películas que él desea ver. Por ejemplo, si se desea ver una comedia se puede filtrar por esa categoría. O si se está buscando algo de terror le aparecerían las películas de esa variedad.

Un tercer filtro seleccionado es por el horario de las funciones. El filtro más innovador es poder seleccionar las películas según el horario de las funciones. Tanto sea entre horario de comienzo como horario de fin. Por ejemplo, si quiero obtener todas aquellas películas que comiencen después de las 20 horas, si utilizamos el filtro con esa hora, se obtiene una lista de todas las películas que comienzan después de las 20 horas.

Vale agregar que todos estos filtros pueden ser utilizados individualmente o pueden ser utilizados en conjunto dependiendo de las necesidades del usuario. De esta manera, se pueden filtrar por género y por horario. Un ejemplo sería filtrar por películas infantiles que tengan funciones después de las 18 horas.

Si bien se seleccionaron estos filtros se pueden agregar más filtros, como ser el año o el país. Se pueden agregar o quitar filtros dependiendo del usuario y las necesidades del mismo. Pero se explica con más detalle las extensiones posibles que se le podrían realizar al proyecto en la sección *Trabajo a futuro.*

3. Conclusiones

Si bien existen muchos sitios con información de películas, tanto sea de los propios cines, como otros secundarios, algunos permiten realizar filtros sobre las películas que los usuarios quieran ver. Sin embargo, con nuestra propuesta creemos que se puede lograr una unificación entre todos los filtros que el usuario quisiera utilizar. Mejorar la interacción con usuario fue nuestro principal objetivo.

Luego de finalizado el proyecto podemos concluir que se cumplieron los objetivos propuesto. Enfocándonos en el aprendizaje de técnicas de web scrapping se logró realizar un proyecto muy completo.

Debido a que este proyecto fue seleccionado a partir de experiencias personales se logró un buen prototipo que cumplía las expectativas. El mismo permite al usuario tener un sitio con información actualizada y unificada de diferentes fuentes. Y puede ser de aún más utilidad si se avanza en el trabajo a futuro mencionado en el punto siguiente.

a. Trabajo a futuro

Se evaluó la posibilidad de extender el proyecto para que abarque más que solo los cines, como por ejemplo, extender el proyecto a teatros o incluso para poder cubrir otro tipo de eventos desde conciertos hasta festivales.

O inclusive con el proyecto existente, si los propios cines proporcionaran sus propias api para poder extraer la información con cierto criterio establecido se podría realizar desde las propias páginas de los complejos. Obteniendo así la información de primera mano y a medida que cada cine actualice su información se actualizaría en la página sin tener que esperar a que la pagina que se utiliza hoy en día se actualice. Lo cual genera más confianza en los datos y es más dinámico y rápido a cambios.

Otra extensión posible es la creación de una aplicación móvil para poder realizar las búsquedas desde un dispositivo móvil.

Otra extensión que se le puede realizar al prototipo presentado es la adicción de más filtros. El proyecto se basa en la atracción de los filtros, es por esto que se cree que se podría profundizar más en este aspecto. Un primer filtro que se le podría agregar es el geo localizador y en base a eso sugerir cuales son los complejos más cercanos. Este filtro recomendaría salas, y en definitiva sus funciones, según la locación que se marque o el GPS, si es que el usuario está en un dispositivo móvil. Otro filtro que se podría agregar es según el país de origen o el año de la película. Por otro lado, también se podría realizar un filtro por actor, actrices o inclusive el director. De esta manera uno podría escribir *Jennifer Lawrence* y solo se mostrarían las películas en la cual la actriz participa.

4. Referencias - Bibliografía

[1] www.cartelera.com.uy

[2] https://angularjs.org

[3] https://jsoup.org

[4] https://docs.oracle.com/javase/tutorial/jdbc/