Methods and tools for the abstraction of object models in web content by end users *

Alex Tacuri[1,2][0000−0003−3159−5556]

[1] Escuela Superior Politécnica de Chimborazo. Panamericana Sur km 1 y 1/2 . Riobamba. Ecuador
[2] LIFIA, Facultad de Informática, UNLP. La Plata.
atacuri@espoch.edu.ec

Abstract. The main idea is to organize the data exposed in Web site interfaces so that they can be processed or used efficiently, either by a person or automated by a machine through an algorithm. The structure or model that will be used in this approach is that of a graph, where each node will be an object abstracted from the DOM. From the point of view of external structures, currently the information on the web is very little combinable, to be able to establish connections or think about object relationships, which leads to not being able to process information effectively and efficiently, In addition, automating a task becomes complex, so providing a user with a model in which they can integrate this data and information will greatly alleviate this lack of integration.

Keywords. Web search, End-user Programming, Web augmentation.

* Universidad Nacional de La Plata

1 Introducción/Motivación

La web actualmente presenta un sinnúmero de sitios web que presentan información muy importante sobre temas de actualidad, pero si reflexionamos un momento, muchos de estos hablan de un mismo tema, con particularidades de cada uno, que si los integramos podemos alcanzar una mejor compresión del mismo; si bien es cierto, esto se lo realiza de manera inconsciente, puesto que, al momento de obtener lo mejor de un sitio, procedemos a buscar en la web otra información relacionada para seguir enriqueciendo el contenido. El objetivo central de este enfoque es primeramente obtener datos de sitios web por medio de Search APIs, mismas que crean una estructura de cómo acceder al contenido de un sitio en específico, para posteriormente integrarlo en un modelo de datos relacionados, con el fin de ofrecer el resultado de

esta integración a través del mismo navegador por medio de una extensión web; todo esto se realiza al momento de invocar a esta capa de servicio. Además recalcar que todo este procedimiento puede ser realizado por usuarios sin habilidades de programación.

2 Estado del Arte

Pensando en la integración y reutilización de información, existen modelos de desarrollo de aplicaciones que se fundamentan en el proceso de concepción orientado a objetos, uno de ellos es el modelo OOHDM (del inglés, «Object Oriented Hypermedia Design Methodology») mismo que en su primera fase, se debe realizar un diseño conceptual de la aplicación hipermedia, con objetos de dominio o clases y relaciones entre ellos que posteriormente pasan a un diseño de navegación, diseño de interfaces e implantación [1]. Otra metodología es WebML, que de igual forma en su primera fase se fundamenta en la creación de un modelo conceptual de objetos que sirve para la navegación y composición de la aplicación, para su posterior presentación [2]. Desde la perspectiva de que un usuario final pueda construir estructuras externas, sería ideal contar con un modelo de objetos que permita basarse en la extracción, vinculación de información de manera reutilizable. Para realizar este modelo podemos aprovechar que el desarrollo de aplicaciones web se basan fundamentalmente en un diseño basado en objetos, los cuales tienen, como fin, aliviar la complejidad y la reutilización de los mismos, que luego de un proceso de integración o composición, llegan a presentarse al usuario final como instancias de los objetos mediante una interfaz de usuario [3]. Sirviéndonos de esta funcionalidad, y dado que estas solicitudes son finalmente presentadas al usuario de alguna manera en la UI, se puede crear un modelo de objetos en el cliente mediante la extracción de la información de ellas, pensando en la reutilización de estos en diversas aplicaciones.

Una vez que el usuario a abstraído información de la web, se propone crear un editor de modelo de objetos, donde se podrán establecer propiedades y relaciones entre ellos, mismos que son provenientes de distintas páginas web a través de web scraping y automatización de la navegación, pero definido a nivel de objetos del dominio. Con este proceso se alcanza la integración e interoperabilidad de las aplicaciones en una capa creada por el usuario final, mismo que podrá interactuar con el modelo, a través de un lenguaje de consultas y una interfaz interactiva que permita la afinación del modelo, y el acceso a objetos e instancias de sitios que hayan sido agregados. Para la elaboración de modelos de objetos web se piensa utilizar técnicas de abstracción de contenidos web con web scraping y aumentación web en el lado del cliente para la definición de editores visuales. Ahora veamos cómo se encuentran estas técnicas en los últimos años; dentro de la WA han aparecido varios interesantes enfoques para considerar. En este año (2020), Hertel et al. presentaron un enfoque donde incorporaron a todos los involucrados en un sitio web a que puedan participar en las actividades de rediseño web con WA y Desarrollo por Usuarios Finales, del inglés «End-User Development» [4]. Gonzales et al. expusieron otro enfoque donde crearon un framework para facilitar el acceso de los usuarios con discapacidad visual dentro del navegador apoyando la accesibilidad web de los sitios [5]. Adicionalmente el mismo autor et al. desarrollaron una api para que los programadores que no están

familiarizados con SPARQL puedan hacer uso de Linked Open Data y puedan acceder de forma más fácil a esta red de información [6]. Sottile et al. (2019), presenta la creación de Aumentos de Web Semánticos [7] y Firmenich et al. desarrollaron una plataforma para interfaces de usuario distribuidas (DUI, del inglés «Distributed User Interface») del lado del cliente, construida sobre los cimientos del aumento web y el desarrollo del usuario final con la finalidad de sincronizar varias interfaces de usuario final [8]. Fernández et al. presentaron a "Logikós" un desarrollo que permite que los usuarios tengan una herramienta para la toma de decisiones, de múltiples criterios en diferentes sitios web [9]. Bosetti et al. presentaron una herramienta de visualización para describir las operaciones fundamentales necesarias para visualizar datos semiestructurados en la Web [10]. En el año 2018, Firmenich et al. presentaron un enfoque para definir y hacer evolucionar los requisitos de aumento web utilizando prototipos visuales enriquecidos y descripciones textuales, que se pueden mapear automáticamente en artefactos de software en ejecución. Bosetti et al. desarrollaron el enfoque para la Aumentación Web Móvil (MoWA, del inglés «Movil Web Augmentation»), En el año 2017, Aldalur et al. identificaron a WA como una tecnología prometedora para EUD [11]. En el 2013, Capra et al. presentaron búsquedas subrogantes con imágenes es decir agregando imágenes a una búsqueda para un mejor análisis de búsqueda [12]. En todos estos enfoques, radica la idea de tener algún tipo de abstracción y/o extracción del contenido público en la Web, pero ningún ataca la problemática específica que es poder programar las estructuras externas basadas en modelos de objetos complejos, que van más allá de solo crear una indirección al elemento del DOM concreto.

3 Planteamiento del problema/Contribuciones

Actualmente la extracción de datos de páginas web (*web scraping*) es utilizado con diferentes propósitos; pero un problema que se ha encontrado es que los datos que se obtuvieron de diferentes fuentes no se encuentran integrados al momento de la extracción; proceso que se lo realiza posteriormente a la abstracción del contenido de la web. En este sentido, nuestro enfoque pretende extraer contenido e integrarlo de forma automática pero no solo por usuarios que dominan programación sino también por usuarios finales sin experiencia en desarrollo de sistemas. Para alcanzar este objetivo se pretende elaborar herramientas con aumentación web que permitan abstraer el contenido web (Search APIs), integrar el contenido y formar modelos de objetos que permitan un mejor procesamiento de la información de diferentes fuentes de datos heterogéneas. Adicionalmente se ha logrado crear una capa de servicio en el navegador que permite que otras aplicaciones puedan consumir datos de esta capa, un ejemplo de este consumo es una aplicación que consume datos de un modelo que integra 3 buscadores: Google Escolar, Springer y DBLP; permitiendo ejecutar búsquedas en estas 3 Search APIs e integra los resultados de cada uno.

4 Metodología de Investigación y Enfoque

Dos grupos uno de control y otro experimental para probar las herramientas desarrolladas. El tipo de investigación bajo la cual se va a manejar la investigación es correlacional puesto que vamos a tener dos variables involucradas; la independiente que será el enfoque de desarrollo de modelos web a partir de abstracciones de la web , y la variable dependiente mejorar la integración de información por parte del usuario final. La idea es verificar si la correlación entre estas variables es significativa. La investigación se va a realizar en la Universidad Nacional de la Plata – Argentina y en Escuela Superior Politécnica de Chimborazo – Ecuador.

5 Plan de Evaluación

Dentro del desarrollo de este enfoque se han realizado los siguientes componentes: una herramienta de abstracción de contenidos web basado en webscraping (Search API), un editor de modelos de objetos que permita crear relaciones y configuraciones entre los objetos previamente abstraídos, una consola interactiva que me permita buscar e interactuar con instancias de objetos basados en el modelo desarrollado por el usuario final utilizando tags semánticos y propiedades, definiendo para esto un lenguaje de consultas sobre el modelo y una aplicación de aumentación web tipo mashup que consume la capa de servicio que se encuentra en el browser como extensión web y permite realizar la consulta del modelo.

Actualmente se está trabajando en realizar el experimento con usuarios finales con y sin experiencia en programación y poder mostrar de forma cuantitativa la validez de las herramientas desarrollados y el enfoque como tal. Posteriormente se pretende utilizar técnicas de machine learning para soportar el proceso de construcción de modelos de objetos o tareas autómatas durante el desarrollo del enfoque.

6 Resultados Preliminares o Intermedios

Se público el artículo “ANDES: An approach to embed search services on the Web browser” en “Computer Standards & Interfaces” sobre Search APIs, indexada en ScienceDirect. DOI: https://doi.org/10.1016/j.csi.2022.103633

Actualmente se desarrolló la capa de servicios que se ejecuta como extensión del browser y es consumida por otra aplicación para hacer consultas, para el experimento vamos a utilizar 3 Search APIs: Springer, DBLP y Google Scholar.

7 Conclusiones y Lecciones Aprendidas

El tiempo en integrar información de varios sitios de un mismo contenido se reduce significativamente logrando mejorar la productividad al momento de buscar información.

8 Etapa Doctoral

Middle.

References

1. J. Molina-Ríos and N. Pedreira-Souto, “Comparison of development methodologies in web applications,” Information and Software Technology, vol. 119. Elsevier B.V., Mar-2020.
2. D. Granada, J. M. Vara, M. Brambilla, V. Bollati, and E. Marcos, “Analysing the cognitive effectiveness of the WebML visual notation,” Softw. Syst. Model., vol. 16, no. 1, pp. 195–227, Feb. 2017.
3. A. Abouzahra, A. Sabraoui, and K. Afdel, “Model composition in Model Driven Engineering: A systematic literature review,” Information and Software Technology, vol. 125. Elsevier B.V., p. 106316, Sep-2020.
4. H. Hertel, A. Dittmar, and D. Linke, “Meta-level support for facilitating participation in website (re-)design activities,” in Proceedings of the 12th ACM SIGCHI Symposium on Engineering Interactive Computing Systems, 2020, pp. 1–6.
5. C. González-Mora, I. Garrigós, S. Casteleyn, and S. Firmenich, “A web augmentation framework for accessibility based on voice interaction,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020, vol. 12128 LNCS, pp. 547–550.
6. C. González-Mora, I. Garrigós, and J. Zubcoff, “An apification approach to facilitate the access and reuse of open data,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020, vol. 12128 LNCS, pp. 512–518.
7. C. Sottile, S. Firmenich, and D. Torres, “An End-User Semantic Web Augmentation Tool,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, vol. 11553 LNCS, pp. 239–243.
8. S. Firmenich, G. Bosetti, G. Rossi, M. Winckler, and J. M. Corletto, “Distributed Web browsing: supporting frequent uses and opportunistic requirements,” Univers. Access Inf. Soc., vol. 18, no. 4, pp. 771–784, Nov. 2019.
9. A. Fernández, G. Bosetti, S. Firmenich, and P. Zaraté, “Logikós: Augmenting the Web with Multi-criteria Decision Support,” in Lecture Notes in Business Information Processing, 2019, vol. 348, pp. 123–135.
10. G. Bosetti, S. Firmenich, M. Winckler, G. Rossi, U. C. Fandos, and E. EgyedZsigmond, “An end-user pipeline for scraping and visualizing semi-structured data over the web,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, vol. 11496 LNCS, pp. 223–237.
11. I. Aldalur, M. Winckler, O. Díaz, and P. A. Palanque, “Web Augmentation as a Promising Technology for End User Development,” in New Perspectives in EndUser Development., F. Paternò and V. Wulf, Eds. Springer International Publishing, 2017, pp. 433–459.

12. R. Capra, J. Arguello, and F. Scholer, “Augmenting web search surrogates with images,” in 22nd ACM International Conference on Information and Knowledge Management, CIKM’13, San Francisco, CA, USA, October 27 - November 1, 2013, 2013, pp. 399–408.