Desarrollo de una Herramienta Interactiva Para la Construcción de un ”Ground Truth”de Segmentaciones de Páginas Web

1nd Andrés Sanoja
Escuela de Computación, Universidad Central de Venezuela
Caracas, Venezuela
andres.sanoja@ciens.ucv.ve

2st Jean García
Escuela de Computación, Universidad Central de Venezuela
Caracas, Venezuela
jean.garcia@ciens.ucv.ve

***Resumen*—En el presente paper describimos los resultados de nuestra investigación, donde se evidencia la importancia de la evaluación de algoritmos de segmentación. La finalidad de nuestro trabajo es la obtención de una ”ground truth”(base de la verdad) de segmentaciones manuales sobre una página Web para la posterior obtención de “la mejor segmentación”, la cual puede ser usada más adelante para la evaluación del algoritmo de segmentación.**

***Index Terms*—segmentación, página web, segmentación manual, ground truth**

I. Introducción

La página Web es un documento digital de información accesible mediante un navegador de Internet, esta información se presenta generalmente en formato HTML, está compuesta por un conjunto de elementos ordenados en una estructura de árbol (el árbol DOM), generado por el navegador a partir del código fuente HTML [1]. La segmentación es definida por la RAE (Real Academia Española) como el acto o consecuencia de segmentar (i.e dividir, formar segmentos o porciones) [2]. En el caso del presente trabajo investigativo, se habla de la segmentación de una página Web, la cual consiste en dividir dicha página Web en fragmentos coherentes (i.e cada fragmento debe tener un sentido para un usuario) llamados bloques. Cada bloque representa distintos elementos de información en la página. Los algoritmos de segmentación son aplicados en diferentes áreas como: Procesos de SEO (Search Engine Optimization, La segmentación de la página Web permite llevar a cabo un análisis del contenido de la página para que ésta pueda ser calificada y ubicada en un ranking.), Migración de Formatos (La segmentación de la página Web en formato HTML4 permite la creación de una nueva página Web usando el formato HTML5 al poder identificar los segmentos de la página anterior), Archivamiento de la Web (Web Archiving, La segmentación de la página Web permite comparar dos versiones de la misma página Web, la versión que actualmente se tiene almacenada y la versión que se planea almacenar, y encontrar las diferencias entre ellas, esto permite detectar si resulta eficiente descargar y almacenar la nueva versión), Bloqueo de Contenido (Content Blocking, La segmentación de la página Web permite poder identificar los segmentos dentro de la página que poseen contenido no deseado para que pueda ser bloqueado). En cada una de estas áreas se desea no solo un algoritmo de segmentación que pueda identificar correctamente la distribución de las segmentos dentro de una página Web, sino aquel que se ajuste a las necesidades de cada área, dependiendo de si se desea un algoritmo de segmentación genérico o un algoritmo de segmentación específico (entendiendo que mientras más genérico sea el algoritmo podrá ser usado en un amplio espectro de páginas Web pero irá reduciendo su certitud. Para poder identificar qué tan genérico o específico resulta el algoritmo de segmentación se debe realizar una evaluación sobre dicho algoritmo, esta evaluación se realiza al comparar la segmentación obtenida con el algoritmo contra una segmentacion denominada como ”la mejor segmentación”, hasta ahora ”la mejor segmentación”se encuentra basada en la segmentación manual realizada por el usuario, ya que no conocemos de otro mecanismo preciso que no sea el manual, además, la correcta distribución de los segmentos dentro de una página es algo que puede resultar subjetivo dependiendo de la visualización que posea el usuario [3]. Ambas segmentaciones deben estar ajustadas a una granularidad, la cual define el tamaño general de los bloques dentro de una segmentación. El presente trabajo facilita el llevar a cabo la evaluación de un algoritmo de segmentación al presentar una herramienta que permite a los usuarios realizar segmentaciones manuales de páginas Web, de forma cómoda, rápida e interactiva. Las segmentaciones generadas a partir de esta herramienta permiten la obtención de una ground truth, la cual representa una base de información obtenida mediante observación. Por medio de esta base se obtiene un elemento vital para la evaluación del algoritmo de segmentación, “la mejor segmentación”.

Contribución: En este paper describimos el funcionamiento y los modelos utilizados en el desarrollo de la herramienta de segmentacion manual (MoB Extension) utilizada porlos usuarios para segmentar manualmente una página Web y enviar dichos resultados al MoB API, Se describe también los procesos que utiliza el MoB API para conformar y analizar el ground truth de dicha página Web para obtener la mejor segmentación. También se describe la interfaz Web

(MoB Repository) que el usuario puede usar para observar los resultados de su segmentación manual así como también observar la mejor segmentación elegida por el sistema.

Organización: En la sección 2 presentamos los trabajos de referencias en los cuales se enmarca este trabajo. En la sección 3 describimos la herramienta de segmentación manual, el MoB Extension. En la sección 4 presentamos el MoB API y el MoB Repository, su funcionamiento y finalidad en este sistema. En la sección 5 describimos los resultados obtenidos de los experimentos realizados. En la sección 6 realizamos la conclusiones y trabajos a futuro.

II. Trabajos de Referencia

Este trabajo se enmarca en la evaluación de la segmentación de páginas Web. Es la continuación de un trabajo futuro planteado en el trabajo realizado por Andrés Sanoja y Stéphane Gançarski publicado en Agosto del 2017: "Migrating Web Archives from HTML4 to HTML5: A Block-Based Approach and Its Evaluation". En la investigación " Block-based Migration from HTML4 Standard to HTML5 Standard in the Context of Web Archives" [4], se presentan dos herramientas mediante lascuales es posible realizar el proceso de migración de formato, estas son: BoM y MoB (Manual-design-Of-Blocks), BoM representa el algoritmo de segmentación el cual segmenta una página Web de forma automática, y el MoB es la herramienta de segmentación manual que nace de la necesidad de asegurar un ground truth para apoyar la evaluación de los algoritmos de segmentación, fue desarrollado un prototipo de baja fidelidad, usado como una extension para el explorador Chrome, permite a usuarios expertos realizar segmentaciones manuales sobre una página Web, un ejemplo se visualiza en la Fig.1

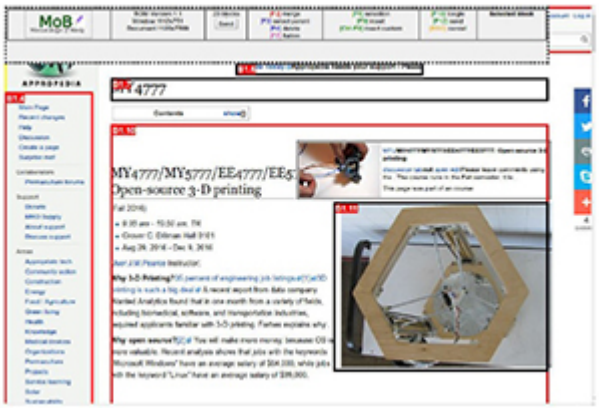

Figura 1. Prototipo de MoB

Los usuarios crean bloques dependiendo de los elementos Web y sus respectivas jerarquías. Se obtiene un grafo de bloques (dadas las jerarquías) o simplemente una segmentación plana (solo bloques terminales), es decir, permite la segmentación en multiples niveles. Ambas segmentaciones producen un documento XML el cual representa el árbol de segmentación basado en el árbol DOM de la página segmentada. Produce también un conjunto de rectángulos presentados de manera visual.

En la Fig.1 se evidencia el hecho de que MoB posee un panel en la parte superior donde muestra la leyenda de los comandos a usar para realizar la segmentación manual. Se le presenta al usuario una segmentación realizada previamente por BoM. Se le propone al usuario aceptar o modificar la segmentación propuesta. Si el usuario desea modificar la segmentación, por ejemplo agregar un bloque, debe presionar F9 y hacer click en el elemento que desea segmentar. MoB mostrará un mensaje con la lista de posibles elementos que se encuentra debajo del click. Este proceso es reflejado en la Fig.2.

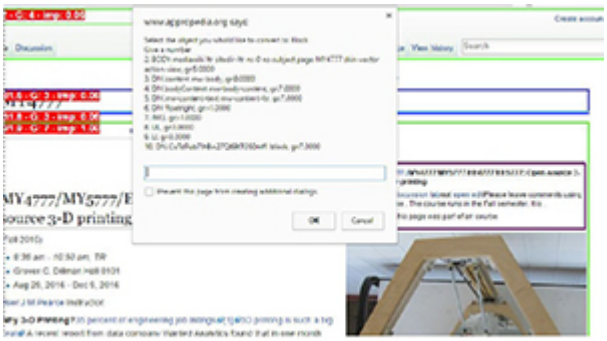

Figura 2. Creación de bloque con el Prototipo MoB

Se debe introducir el número del elemento y presionar el botón de "ok", una vez hecho esto se resaltará dicho elemento indicando que se ha segmentado. Este proceso resulta tedioso, complicado y propenso a errores, incluso para usuarios expertos, por esta razón el presente paper presenta una mejora en la usabilidad de MoB. Por ejemplo, resaltar tentativamente el elemento que se desea segmentar permitiéndole al usuario una retroalimentación activa y se crea también un sistema de puntajes, el cual motiva a los usuarios a realizar segmentaciones. Todas las nuevas y mejoradas características de la herramienta MoB se describen en la siguiente sección.

III. MoB Extension

La extensión de MoB es la herramienta de segmentación manual desarrollada en nuestra investigación con el lenguaje de programación Javascript, el lenguaje de marcado HTML5 y las reglas de estilo CSS3, está basada en el prototipo de MoB creado por Andrés Sanoja y Stéphane Gançarski. Actualmente está desarrollada para ser usada en navegadores de Chrome y Chromium como una extensión.

III-A. Funcionalidades de la Extensión MoB

- **Consultar información:** posee una sección de información donde el usuario puede aprender más de la herramienta y sus diferentes acciones.
- **Cambiar idioma:** permite al usuario cambiar el idioma de la interfaz escogiendo entre: ingles, frances o español.
- **Consultar puntuaciones:** permite a un usuario autenticado en el sistema consultar sus puntuaciones personales

sobre una determinada página Web o las puntuaciones globales.
- **Registro:** permite al usuario registrarse dentro del sistema MoB.
- **Autenticacion:** una vez el usuario esta registrado, el sistema le permite autenticarse mediante un inicio de sesión.
- **Segmentación manual:** el sistema le permite al usuario identificado inicializar la herramienta de segmentación manual.

III-B. Acciones de la Herramienta de Segmentación

Las funcionalidades anteriormente descritas forman parte de la extensión MoB en general, se considera pertinente describir a continuación las acciones que ofrece la herramienta de segmentación de la extensión MoB. Al inicializar la herramienta de segmentación manual, aparecerá un menu con las siquientes acciones:

- **Agregar nuevo bloque:** permite agregar un nuevo bloque a la segmentación, al ser seleccionada el usuario puede recorrer los elementos del DOM con el mouse y estos serán iluminados, una vez el usuario haga click sobre alguno de estos elementos del DOM, se inyectará en el HTML un cuadrado representando el bloque segmentado, dicho elemento posee entre sus atributos los datos del bloque segmentado (etiqueta, ancho, alto, posición en left, posición en top, área), dichos datos son también almacenados dentro de un arreglo en Javascript. Este arreglo es el que es utilizado para realizar las comparaciones necesarias entre los diferentes bloques para llevar a cabo las futuras acciones. Cabe destacar que al realizar esta acción se activa un subproceso que verifica si hay otros bloques dentro del que está a punto de crearse, en cuyo caso se eliminan dichos bloques y se conserva únicamente el nuevo que ha sido creado. En caso de que existan bloques que intercepten mas no se encuentren completamente dentro del nuevo, ocurriría un error de intercepción el cual requerirá de una acción de “cortar”.
- **Eliminar bloque:** al estar esta acción activada, se iluminaran aquellos bloques al que el usuario señale con el mouse, al hacer click sobre alguno de estos bloques se disparará una función de Javascript la cual removerá el elemento DOM que representa el rectángulo del bloque, y dicho bloque será removido del arreglo de bloques.
- **Unir bloques:** al ser seleccionada, esta acción permite al usuario seleccionar dos bloques los cuales se unirán. Esta acción elimina los dos bloques seleccionados y crea uno nuevo que comparte los límites superiores e inferiores máximos de los bloques anteriores. Después se comprueba que no hayan quedado atrapados otros bloques dentro del nuevo, en cuyo caso se eliminarían.
- **Cortar bloques:** esta acción permite seleccionar dos bloques que se están interceptando (A y B) y realizar un corte entre los dos. El orden de selección importa, pues A será el bloque que predominará (se mantendrá intacto) y B será el bloque que se recortará para solucionar la intercepción.
- **Etiquetar bloque:** al ser seleccionada, esta acción le permite al usuario seleccionar cualquier bloque presente en la segmentación, al hacer click sobre un bloque se mostrará una modal con una lista donde el usuario podrá escoger la etiqueta que mejor se adapte al bloque.
- **Seleccionar bloque:** al estar seleccionada, esta acción permite al usuario seleccionar cualquier bloque y obtener una ventana de información con los datos de dicho bloque.
- **Panel de información:** al ser activada, esta acción despliega un panel informativo con metadatos de la segmentación para el usuario, ofrece la opción de cambiar la granularidad de la segmentación y muestra todas las alertas que puede presentar la segmentación.
- **Enviar segmentación**: esta acción activa los procesos necesarios para la recolección de datos y envío hacia el API de MoB, comienza por comprobar el estado de la segmentación en busca de errores o advertencia, en caso de presentar errores, la segmentación no será enviada y los procesos de recolección de datos se cancelan, en caso contrario, se crea una estructura de JSON para enviar todos los datos necesarios al API, para poder obtener todos los datos, se almacena el HTML actual de la página en un string, esto vendría siendo el HTML de la página versión MoB. A la vez se hace una búsqueda entre los elementos DOM de la página para identificar cuales son los elementos DOM que están presente en cada bloque segmentado, esta información se almacena en un arreglo junto con los bloques de la segmentación, a demás de otros metadatos como el título de la página, la dirección URL, la categoría, colección y dimensiones.

III-C. Finding the Granularity Model

Como se mencionó en la sección anterior, para poder realizar la evaluación se requiere de ”la mejor segmentaciónz la segmentación realizada por el algoritmo, ambas con una misma granularidad, se puede ver el nivel de granularidad como el nivel de detalle de la segmentación, a menor granularidad mayor nivel de detalle.

III-D. Factores Influyentes en la Granularidad

- La relación que existe entre el área de los bloques segmentados y el área total del documento, a mayor tamaño del bloque, mayor es el nivel de abstracción de la segmentación, por lo tanto mayor granularidad.
- La cantidad de bloques que posea la segmentación, a menor número de bloques, mayor abstracción, mayor granularidad.
- El nivel de granularidad puede variar inmensamente de un documento a otro, pero la relación se debe mantener, mientras mayor número de bloques, menor debe ser el tamaño de cada uno. A mayor nivel de granularidad se debe limitar el número de bloques permitido, impidiendo a su vez el decremento del área de los mismos.

III-E. Granularity Model

- En cero de granularidad, se puede tener un gran número de bloques (por lo que los bloques pueden llegar a ser muy pequeños). - En granularidad 10 se desea un solo bloque, el documento completo. - En granularidad 9 se desea un máximo general de 3 bloques (header, content y footer). - En el intervalo [1,8] se desea una interpolación de granularidad con los límites anteriormente descritos.

Se hace notación de que G(x) se lee como "Granularidad del documento a un nivel x". Se tiene que a G(0) se pueden tener tantos bloques como se desee, pero en G(1) se establece un número máximo de bloques de 40. Realizando las interpolaciones necesarias sobre los limites expuestos tenemos los siguientes resultados:

G(0) = 40+ bloques, G(1) = 40 bloques, G(2) = 36 bloques, G(3) = 31 bloques. G(4) = 27 bloques, G(5) = 22 bloques, G(6) = 18 bloques, G(7) = 13 bloques, G(8) = 8 bloques, G(9) = 3 bloques, G(10) = 1 bloque.

De esta forma se obtiene el máximo de bloques para cada nivel de granularidad, Como esto representa el número máximo de "pedazos" en los que se puede dividir el documento, lo lógico sería considerar que el área máxima que cada uno de estos bloques puede llegar a alcanzar representa el área mínima que debe poseer los bloques en este nivel de granularidad, ya que pueden existir menos bloques de los esperados, pero no más.

Se puede calcular el área mínima que debe poseer cada bloque en cada una de las granularidades para una página determinada, esto permite poder clasificar los diferentes bloques dentro de una determinada granularidad, por ejemplo, si el bloque posee un área que se encuentra entre el área mínima de la granularidad 4 y la granularidad 5, se dice que el bloque es de granularidad 4.

Lo ideal sería que la segmentación posea únicamente bloques que posean la misma granularidad de la segmentación, sin embargo, ese no siempre resulta el caso y dado que esto puede resultar muy restrictivo, se decide dar un rango de error de una (1) granularidad, es decir, el bloque aún es aceptado si se encuentra una granularidad por debajo de la granularidad de la segmentación, es decir, si el área del bloque es mayor o igual que el área del documento dividido entre el factor de granularidad de la segmentación menos 1. En la (1) se observa el limite inferior del área que debe tener cada bloque dentro de la segmentacion en una determinada granularidad. Abi representa el area de los bloques, donde i=1,2,3,4,... y Adoc representa el area total de la página Web.

$$A_{bi} \geq \frac{A_{doc}}{G(x-1)} \tag{1}$$

III-F. Diferencias entre las dos Versiones de MoB

En la Fig.3 se puede observar que la nueva herramienta posee las siguientes características a diferencia de la antigua versión:

Figura 3. Comparación entre ambas versiones de MoB

- El color de los bloques de la nueva versión refleja el estado de los mismos y no el nivel de la segmentación, pues la segmentación es plana.
- La nueva versión restringe al usuario a realizar una segmentación plana, lo cual se considera deseable, a diferencia de la version anterior la cual es propensa a segmentar en multiples niveles.
- En la nueva versión el panel de información ofrece mayor información sobre la granularidad presente y los bloques, a demás de los posibles errores o advertencias que puedan ocurrir.
- En la nueva versión la caja de herramienta ocupa menos espacio al estar conformada únicamente de metáforas de las acciones, también es semi-transparente para poder observar el contenido que existe detrás.
- En la nueva versión todas las acciones son llevadas a cabo únicamente con el mouse, sin tener que usar el teclado.
- La nueva versión ofrece la acción de "cortar" para poder separar bloques que se intersectan, y la acción "seleccionar" que permite obtener toda la información de un bloque en específico.

IV. MoB API y MoB Repository

El MoB API y el MoB Repository están estrechamente relacionados, ya que, el MoB API no sólo ofrece servicios RESTful sino también actua como backend del MoB Repository. A continuacion se hace una descripción de las características más resaltantes de ambos y su finalidad en el sistema.

IV-A. MoB API

El desarrollo de la API se realizó con el lenguaje de programación Python v.3.5, apoyado con el microframework Flask v.0.12.2. Para la creación de la base de datos que va conectada a la API se utilizó el manejador de base de datos Postgresql v.10.1 junto con un componente llamado Postgis 2.4 para realizar las operaciones entre tablas.

En general, el API de MoB se divide en dos partes: todos los servicios RESTful que pueden ser ofrecidos al MoB Extension o a terceros y todas aquellas funciones que manejan el backend del Repositorio MoB.

A continuación se describen los servicios RESTful que son ofrecidos por la API, mientras que las funciones que manejan el backend del Repositorio de MoB se explicarán en la siguiente subsección. Los servicios ofrecidos son los siguientes:

- **Registrar usuario:** Permite registrar a un usuario en el sistema, para completar el registro se le envía al usuario un link de activación a su correo.
- **Iniciar sesión:** Permite al usuario registrado (y activado) iniciar sesión en el sistema para hacer uso de sus funcionalidades.
- **Cerrar sesión:** Borra las cookies de sesión existentes en el navegador y la sesión existente en el API.
- **Recuperar contraseña:** Permite al usuario recuperar su contraseña en caso de extravío, el sistema envía una combinación aleatoria de caracteres como contraseña temporal dado a que por medidas de seguridad las contraseñas se encuentran encriptadas por hash en la base de datos.
- **Obtener colecciones:** Permite obtener una lista con los nombres de las colecciones y categorías de éstas existentes en la base de datos del sistema.
- **Obtener etiquetas:** Permite obtener una lista con los nombres de las etiquetas existentes en la base de datos del sistema.
- **Obtener puntajes globales:** Permite obtener una lista con los mejores puntajes en cada una de las granularidades de una página Web específica.
- **Obtener puntajes del usuario:** Permite obtener una lista con los puntajes en cada una de las granularidades de una página Web específica para un usuario determinado.
- **Cargar segmentación:** Este es uno de los servicios más importante del API pues representa la base de todo el sistema, permite cargar los resultados de una segmentación a la base de datos (y los datos de la página Web en caso de que sea la primera vez que se segmenta). Formando de esta forma lo que denominamos anteriormente como "Ground Truth".
- **Vista previa de segmentación:** Devuelve un canvas con las figuras y etiquetas de los bloques segmentados para una segmentación en específica.
- **Obtener segmentación en formato JSON:** Retorna un JSON con todos los datos de una segmentación en específica.
- **Obtener segmentación en formato V-PRIMA:** Retorna todos los datos de una segmentación específica en formato V-PRIMA, el formato V-PRIMA consta de un XML donde se especifican los bloques existentes en la segmentación y los links, imágenes y textos existentes dentro de éstos.
- **Obtener segmentación en formato MoB HTML:** Dada una segmentación determinada, retorna un HTML con la información que se capturó momentos antes de enviar la segmentación, es decir, el HTML original de la página Web modificado por la herramienta MoB tras realizar la segmentación.
- **Obtener página Web en formato WARC:** Devuelve la información de una página Web en formato WARC (Web ARChive), el formato WARC permite la concatenación de múltiples objetos de datos o recursos en un solo archivo, de esta forma es utilizado para almacenar la información de páginas Web junto con sus recursos y metadata.

IV-B. Mob Repository

El desarrollo del sitio Web MoB Repository se desarrolló haciendo uso del lenguaje de marcado HTML5 para la estructura de las páginas, las reglas de estilo CSS3 para la apariencia de las mismas, y el lenguaje de programación Javascript junto con un framework del mismo llamado Jquery para el comportamiento de las páginas y control de eventos. Todo esto del lado del cliente, del lado del servidor, como se mencionó antes, está apoyado por el MoB API (Python y Flask) y la base de datos conectada a éste (Postgresql y Postgis).

La finalidad del MoB Repository es ofrecerle a los usuarios del sistema una interfaz para que puedan visualizar las colecciones de segmentaciones manuales almacenadas en la Base de Datos, y observar "las mejores segmentaciones"de cada página Web segmentada. A los usuarios administradores les permite administrar las etiquetas utilizadas en la segmentación, las colecciones y sus categorias, asi como los roles de otros usuarios.

IV-C. Best Segmentation Model

Para poder obtener "la mejor segmentación"de forma confiable, se debe crear, a partir de un pool de segmentaciones (Ground Truth).La mejor segmentación debe ser aquella que comparta la mayor cantidad de similitudes entre las demás segmentaciones, dado que, los usuarios pueden llegar a tener diferentes puntos de vista, pero los puntos de vistas que se comparten aseguran un acuerdo. Mientras más compartida

sea dicha similitud, la definición de la misma se vuelve más verídica, al fin y al cabo la segmentación de la página es definida por cómo el usuario promedio la percibe.

Al llevar a cabo la comparación, los bloques recibirán una serie de puntajes dependiendo de sus similitudes geométricas y semánticas con respecto a los otros bloques, los puntajes de todos los bloques correspondientes a una misma segmentación podrán ser sumados dando la puntuación total de la segmentación, esto a su vez permite asegurar cierta fidelidad a la mejor segmentación a través de su puntaje, el cual no solo refleja si fue la mejor sino también la mejor entre cuantos (mientras mayor sea el puntaje obtenido, mayor es el número de segmentaciones que la respaldan). Varios de estos atributos se nombran durante el proceso de comparación, por lo que aquí se presenta una leyenda de los mismos:

- **Identificador:** Un identificador arbitrario del bloque.
- **Geometría:** La geometría del bloque (área, ubicación, entre otros)
- **Score Geométrico:** La puntuación obtenida por similitudes geométricas.
- **Etiqueta:** La etiqueta que le fue asignada al bloque.
- **Score Semántico:** La puntuación obtenida por similitudes semánticas.

El proceso de la creación de la mejor segmentación consta de las siguientes 3 etapas:

1. **Identificación de los bloques:** principalmente se deben identificar los bloques de la nueva segmentación a ser evaluada (llamemosla segmentación N) con respecto a los ya almacenados en la base de datos de las demás segmentaciones (llamemoslas Oi, donde i=1,2,3...), ya que se desea llevar un control de los bloques similares, es por esto que todos los bloques similares (geométricamente) llevan el mismo identificador arbitrario. Se toman los bloques de la segmentación N y se comparan por coincidencias entre todos los bloques de Oi (con un cierto margen de error), para ello se mide la Distancia Hausdorff [1] entre un bloque de N y todos los bloques de Oi, si la distancia es menor o igual a 30 se considera que ambos bloques son similares y el bloque de N heredaría el identificador que posea el bloque de Oi al que es similar.
2. **Contabilización de puntos:** para complementar el paso anterior, se contabilizan todos los bloques bajo un mismo identificador, para obtener el score geométrico, después, dentro del mismo pool de bloques con el mismo identificador se contabilizan todos los que posean las mismas etiquetas, de esta forma obtener el score semántico.
3. **Creación de la mejor segmentación:** para crear la mejor segmentación se incluyen todos aquellos bloques (uno por cada identificador) cuyo score geométrico sea mayor que el 50 % del número de segmentaciones realizadas, esto garantiza que la mayoría de los usuarios opina que ese bloque debe existir. Dicho bloque poseerá la etiqueta más utilizada para ese bloque, es decir, se busca entre los de un mismo identificador la etiqueta que tenga el mayor score semántico.

[1]La distancia Hausdorff es la mayor de todas las distancias existentes desde un punto en un conjunto hasta el punto más cercano en otro conjunto.

V. Resultados

Para evaluar el sistema desarrollado, se llevaron a cabo dos pruebas de aceptación, una funcional y otra no funcional, las cuales se describirán a continuación.

V-A. Prueba Funcional

Para comprobar la funcionalidad del sistema, se realiza una prueba de caja negra. La cual consiste en comprobar si el sistema se comporta como es esperado en cada una de sus funcionalidades, no interesa la lógica de los procesos, únicamente los datos de entrada y salida (o el estado del sistema inicial y final)Para comprobar la funcionalidad del sistema, se realiza una prueba de caja negra. La cual consiste en comprobar si el sistema se comporta como es esperado en cada una de sus funcionalidades, no interesa la lógica de los procesos, únicamente los datos de entrada y salida (o el estado del sistema inicial y final). Para ese fin se realiza un control sobre todos los casos de uso presentado por la MoB Extension, el MoB API y el MoB Repository, se realiza dicho caso de uso y se comprueba que el resultado sea el esperado, en estas pruebas se evidenció que los trem componentes del sistema (la MoB Extension, la MoB API y el MoB Repository) se comportan como se esperaba.

V-B. Prueba no Funcional

Para las pruebas no funcionales se quiere medir la usabilidad del sistema, específicamente de la herramienta de segmentación manual MoB, para esto se evalúa la reacción se 5 individuos ante el sistema. Se observó las reacciones de los participantes mientras completan 2 objetivos planteados y finalmente se les dió un cuestionario para que responder según su experiencia. Cabe acotar que la herramienta está orientada a usuarios de 13 años en adelante, con inexistentes, bajos, intermedio o avanzados conocimientos en segmentación de páginas Web. Los objetivos fueron:

- Realizar una segmentación manual sobre la página Web: https://wiki.apache.org/httpd/RedirectSSL.
- Visitar el Repositorio de MoB y observar la segmentación realizada.

En conclusiones generales, basándose en las respuestas obtenidas del cuestionario, los comentarios hechos por los participantes y el comportamiento observado de los mismos, se tiene que:

- El sistema en general presenta un aspecto estético agradable para los usuarios.
- La herramienta de segmentación permite a los usuarios inexpertos realizar segmentaciones rápidamente sobre una página Web, de una forma sencilla.
- La navegación general del Repositorio MoB es entendible, sin embargo cuando se debe profundizar, buscar

segmentaciones específicas, el usuario debe invertir un poco de tiempo en entender la lógica de la navegación.

Se puede considerar que el sistema MoB es usable, sin embargo, dada la complejidad del mismo, requiere de un breve periodo de aprendizaje por parte del usuario.

VI. Conclusiones y Trabajos a Futuro

Durante el desarrollo de la herramienta de segmentación manual (Extensión de MoB), se presentó uno de los retos más grandes del proyecto. Se requirió no sólo que cumpliese con su misión de segmentar la página Web, lo cual ya de por sí es retador cuando el mundo del diseño de las páginas Web es muy amplio y se encuentra en constante cambios, pero además se requería presentarle al usuario una herramienta de fácil uso. Por ejemplo, que al alcance de un clic pudiese ser capaz de recortar una sección de la página Web, que pudiese editar la sección creada de diferentes formas e incluso recortar posibles secciones que se intersectan, entre otras funcionalidades, todo de forma cómoda, rápida y entendible para el usuario. Al cabo de las pruebas de caja negra y usabilidad se dio a conocer que el objetivo fue logrado, aunque el usuario debe pasar por un breve periodo de instrucción previo ofrecido por el mismo sistema donde se le explica la tarea que debe realizar, esto se debe a la complejidad inherente que presenta la tarea de realizar una segmentación manual sobre una página Web.

En cuanto al repositorio de MoB, el mayor reto fue la forma de encontrar una forma comprensible de ordenar toda la información sobre las páginas Web y sus segmentaciones, después de las pruebas de caja negra se evidencio que el repositorio funcionaba como se esperaba, aunque en las pruebas de usabilidad se evidenció cierta dificultad por parte del usuario al momento de navegar por el repositorio de MoB, ya que el usuario requería un poco de tiempo para entender la lógica del sistema. Esta información recopilada puede ser tomada en cuenta para las mejoras a futuro que se vayan a realizar sobre el sistema.

Entre los retos presentados en el desarrollo del API de MoB se encontraba: el poder desarrollar todos los servicios pertinentes de una forma modular, presentando estructuras de respuesta que fuesen fácil de comprender y manejar, en especial la estructura de bloques que se debe pasar para cargar los datos de la segmentación, también presentó un reto el análisis de la mejor segmentación, dicho proceso fue desarrollado para ser ejecutado en el API de MoB, como un hilo paralelo para no mantener esperando al cliente, este era uno de los retos presentados para el análisis, ya que se debe realizar dicho análisis cada vez que se carga una nueva segmentación en el sistema, otro de los retos del análisis fue la forma de comparar los diferentes bloques de forma rápida y eficaz, gracias a Postgis con sus objetos Geométricos y funciones asociadas a estos, este reto pudo ser superado sin muchos inconvenientes.

Este proyecto comenzó con la idea de desarrollar una herramienta para ayudar en el desarrollo de otros proyectos. Sin embargo, a lo largo del desarrollo se fueron extendiendo las funcionalidades principales y añadiendo funcionalidades adicionales a los elementos principales del sistema, reforzando la funcionalidad y permitiendo la evolución del mismo. Como resultado tenemos un sistema bastante completo en donde se pueden realizar segmentaciones manuales, incluir segmentaciones hechas por algoritmos y ser almacenadas. La información de las segmentaciones puede ser mostrada a través de una interfaz Web (Repositorio MoB), y en el caso de las segmentaciones manuales, pueden ser analizadas y obtener ese elemento que representa la razón principal por la cual se realiza este trabajo investigativo, la mejor segmentación.

Se considera que en la realización de este trabajo se completaron exitosamente los objetivos planteados e incluso se dio un paso extra, sin embargo esto no significa que este trabajo representa la solución definitiva para la problemática expuesta, se espera que el presente trabajo especial pueda servir de base o inspiración para más trabajos relacionados que experimenten con otros enfoques.

VI-A. Contribución

Como se ha mencionado a lo largo de esta investigación, el presente trabajo representa en sí un elemento muy importante en un sistema mucho más grande. El resultado del sistema desarrollado es aquella segmentación llamada “la mejor segmentación”. Esta segmentación representa la segmentación que será usada como parámetro de evaluación para evaluar los algoritmos de segmentación. Es importante mencionar que este trabajo está enmarcado en un proyecto investigativo mucho mayor, existe un trabajo en progreso realizado por Brayhan Villalba el cual tomará el resultado de este trabajo para desarrollar el proceso de evaluación del algoritmo de segmentación.

VI-B. Trabajos Futuros

La inclusión de otros formatos de exportación para las segmentaciones sería una buena actualización.

Según los resultados arrojados por la prueba de usabilidad, el Repositorio MoB presenta una navegación que puede resultar algo confusa para algunos usuarios, es por eso que se propone como un trabajo a futuro la implementación de una navegación más intuitiva.

A demás, es importante destacar que la extensión actual funciona únicamente para el navegador Chrome/Chromium es por eso que se plantea un trabajo a futuro donde se adapte dicha extensión para funcionar en una mayor variedad de navegadores (Firefox, Safari, Opera, entre otros).

Referencias

[1] W3C. (2005). Document Object Model (DOM). Consultado en Marzo-2017. Recuperado de: https://goo.gl/9Xwtqg

[2] Real Academia Española. (2005) . Diccionario Panhispánico de Dudas. Consultado en Abril-2017. Recuperado de: https://goo.gl/SBr586

[3] Cai D, Shipeng Y, Ji-Rong W, Wei-Ying M. (2003). VIPS: a Vision-based Page Segmentation Algorithm. Consultado en Marzo-2017. Recuperado de: https://goo.gl/1FNChD

[4] Sanoja A., Gançarski S. (2017). Migrating Web Archives from HTML4 to HTML5: A Block-Based Approach and Its Evaluation. Consultado en Abril-2017. Recuperado de: https://goo.gl/8imYoy