DETECCIÓN DE PATRONES DE ACCESIBILIDAD EN EL DESARROLLO DE PROYECTOS WEB

Villanueva, J. [(p)]; Rodríguez, V.; Ortega, F.; Mijares, A.

Abstract

The use of accessibility requirements in the development of web projects is required depending on the organization to which the project is developed.

This paper seeks to find patterns of good and bad practices in the development of websites based on the analysis with data mining techniques and the checks carried out using a tool developed for the purpose by CTIC.

For this purpose, data set are taken from 28 different sites to conduct the studies. Based on the methodologies and techniques of data mining, this study aims to identify check groups that allow clustering behaviors of the sites sampled in relation to accessibility so it is possible to detect what checks characterize web sites with good or bad accessibility practices.

Keywords: Data mining, web accessibility projects, information systems.

Resumen

La utilización de criterios de accesibilidad en el desarrollo de proyectos Web es exigible en función del organismo para el que se desarrolle el producto.

En el presente trabajo se pretende encontrar patrones de buenas o malas prácticas en el desarrollo de sitios web a partir del análisis con técnicas de DataMining y de las verificaciones realizadas mediante una herramienta desarrollada a tal efecto por CTIC.

Para ello se parte de una serie de datos procedentes de 28 sitios web diferentes sobre las que realizar los estudios. Tomando como base los estudios, metodologías y técnicas de minería de datos, se pretende conseguir identificar grupos de verificaciones que permitan agrupar el comportamiento de los sitios muestreados en relación a la accesibilidad, con lo cual se pueden detectar que verificaciones caracterizan los sitios web con buenas o malas prácticas de accesibilidad.

Palabras clave: Minería de datos, proyectos de accesibilidad web, sistemas de información.

1. Introducción

Este trabajo surge de la inquietud por parte de la Fundación CTIC (Centro Tecnológico de la Información y de la Comunicación) de analizar un conjunto de datos procedente de una potente herramienta para el análisis de la accesibilidad de sitios web, como es TAW (Test de Accesibilidad Web). Esta herramienta es capaz de analizar y comprobar el nivel de accesibilidad alcanzado en el diseño y desarrollo de páginas Web, realizando sobre cada una de ellas un conjunto de verificaciones que han de satisfacer las pautas de accesibilidad definidas en las reglas desarrolladas por la Iniciativa de Accesibilidad Web (WAI)[1], perteneciente al World Wide Web Consortium (W3C). Estas recomendaciones, denominadas Pautas de Accesibilidad al Contenido Web 1.0 (WCAG 1.0), son normas aceptadas universalmente.

Demostrada la viabilidad de la utilización de las técnicas de minería de datos sobre estos tipos de problemas [2], se procede a contactar con esta organización que dispone de información fiable y de calidad contrastada para comenzar la realización del estudio propuesto en este articulo.

En este caso se trata de un proyecto de minería de datos por lo cual, se plantea la utilización de la metodología CRISP-DM [3] la cual estructura el ciclo de vida de un proyecto de minería de datos en seis fases, las cuales van desde las fases en las que se fijan los objetivos hasta las fases en las que se implanta el modelo resultado del proyecto.

2. Definición de objetivos y conjunto de datos

Siguiendo las fases marcadas por la metodología CRISP_DM, la primera de las fases tiene como propósito analizar el problema y definir los objetivos. En este trabajo se definen como objetivo encontrar los criterios de verificación que agrupa las páginas web que se analizaran, con el fin de identificar cuáles son las verificaciones que caracterizan cada uno de los grupos.

Una vez realizado el análisis del problema, con el fin de convertirlos en objetivos, se realiza la fase del análisis de datos la cual comprende la recolección inicial de datos, en orden a que sea posible establecer un primer contacto con el problema, identificando la calidad de los datos. Para ello se recoge la información que genera el programa TAW con las verificaciones que no genera conformidad respecto a la norma WAI.

Se parte de más de cuatro millones de registros con las verificaciones realizadas a cada una de las páginas que contienen los sitios Web del estudio. Los cuales se han de transformar para convertir en una matriz en la que cada verificación sea una columna. Se dispone de un conjunto de 67 verificaciones sobre las que se realizará el estudio. Cada verificación puede tomar tres valores, dos de ellos asociados a si pasa o no pasa la verificación y el tercero que es duda puesto que por criterios automáticos o se pude garantizar si pasa o no la verificación.

3. Exploración del conjunto de datos

Realizando un análisis inicial de la muestra de datos obtenemos la distribución de las páginas web agrupadas por sitio Web. Se puede ver, que los sitios Web tienen como media un 4% del total de las páginas.

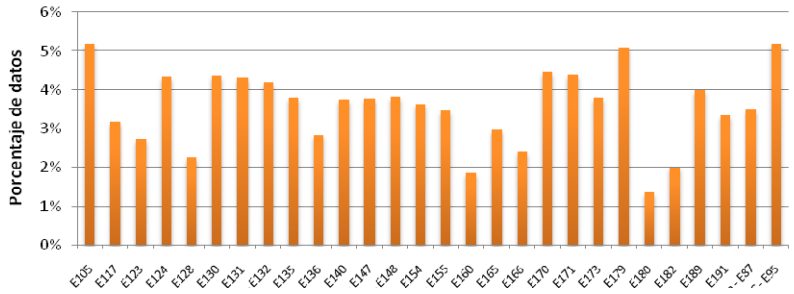


Figura 1. Distribución de las muestras tomadas en cada sitio Web.

Como primera aproximación se ha realizado un estudio de la importancia relativa de cada una de las verificaciones sobre cada uno de los sitios Web, puesto que es uno de los elementos significativos del conjunto de datos, en relación a la accesibilidad.

Para ello, se han entrenado 28 modelos con la técnica MARS [4], los cuales, se les plantea como objetivo la presencia o no de cada uno de los sitios Web y sobre estos modelos se analiza la importancia de cada una de las verificaciones. De esta manera, tendremos una aproximación a las verificaciones que caracterizan cada sitio web, y por tanto, las verificaciones que no son significativas se eliminan.

Para este estudio se ordenan las verificaciones atendiendo a tres criterios:

- Suma de la importancia obtenida por cada una de las verificaciones en cada modelo.
- Número de sitios web en los que la verificación aparece.
- Número de sitios Web en los que la verificación aparece con una importancia relativa superior a un 20%.

Los criterios de clasificación se han categorizado con cuatro valores, que son: “Alta”, “Relativa”, “Poca” y “Nula”.

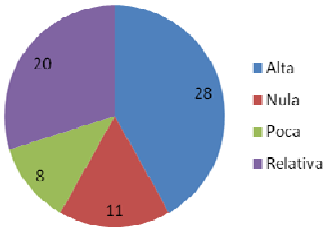


Figura 2. Distribución calidad de las verificaciones.

Que una verificación tenga una importancia “Nula” se refiere a que esta no aporta información a la identificación del sitio Web (esto puede ser porque esa verificación sea constante, puesto que, si una verificación siempre da como resultado “Pasa” no es significativa para la extracción de características de un sitio Web respecto al resto). Por tanto, se han categorizado como verificaciones de importancia “Nula” las que no son necesarias para generar un modelo asociado al sitio Web.

Se categoriza como “Poca” las verificaciones que aparecen en algún modelo, pero tienen una importancia relativa inferior al 20% en ese modelo.

Como resultado de este proceso se procede a eliminar del conjunto de datos las 19 verificaciones etiquetadas como “Nula” y “Poca”.

Por otro lado, después de distintos análisis realizados con el conjunto de datos inicial, se ha observado que existen muchas concurrencias, es decir, páginas con el mismo conjunto de valores para las verificaciones dentro de un mismo sitio Web, con lo que tenemos mucha información redundante.

Se ha optado por eliminar las páginas repetidas dentro de cada sitio Web (dejar un caso, dentro de cada sitio Web en el que aparece), debido a que si eliminamos la totalidad de las páginas repetidas sin tener en cuenta el sitio Web al que pertenecen, esto sólo disminuiría el conjunto de datos en un 0,3% más, con lo que al final se ha considerado que disponer del sitio Web como unidad que englobe las páginas puede aportar información.

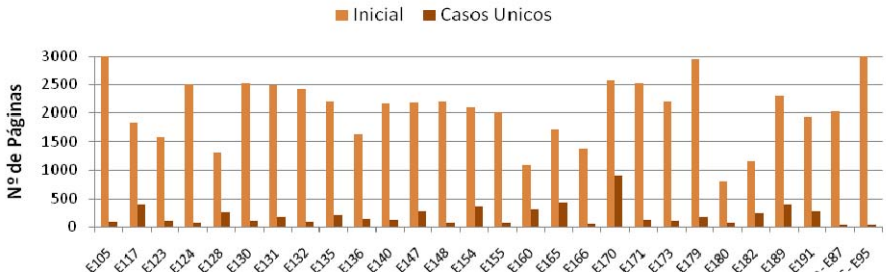


Figura 3. Frecuencia de presencia de datos por cada sitio Web (datos reducidos).

En la figura anterior se muestra una gráfica comparativa del conjunto de datos inicial y del conjunto de datos que sólo conserva una página con el mismo conjunto de valores para las verificaciones, dentro de un mismo sitio Web. Con ello, conseguimos que el conjunto de datos final quede reducido al 9,7% de conjunto de datos inicial.

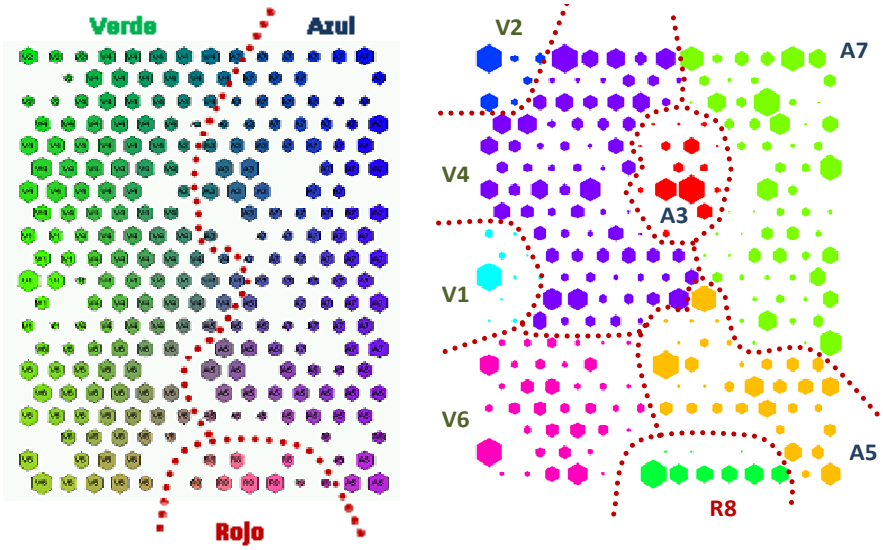
Como se puede ver hay sitios web que contienen un conjunto de páginas más homogéneas en relación a las verificaciones de accesibilidad y por lo tanto se reduce considerablemente el número de páginas que forman el sitio Web.

Por el contrario, también hay sitios web con un comportamiento más heterogéneo con lo que se recude mucho menos el número de páginas de la muestra por sitio Web.

4. Análisis y búsqueda de agrupaciones

Una vez filtrado y creado un nuevo conjunto de datos que contienen sólo las verificaciones significativas (48) y los casos únicos (páginas no duplicadas) por sitio Web, se procede a la representación del mapa SOM [5].

En este caso, se ha utilizado el color como indicador de distancia entre las celdas, es decir, las celdas que tengan un color similar se encuentran cerca. De esta manera, se puede observar que existen tres grandes grupos como se muestra en la figura siguiente, a los que se hará referencia en el resto del artículo con las etiquetas de Verde, Azul y Rojo.



(a) (b)

Figura 4. (a) Mapa SOM con conjunto de datos reducido. (b) Mapa SOM con conjunto de datos reducido, detallado por agrupaciones

Si realizamos un análisis más detallado sobre los grupos Verde y Azul de la Figura 4 (a) obtenemos la clasificación que se puede ver en la Figura 4 (b), en la que se ha utilizado el color para separar los grupos y subgrupos.

El tamaño de la celda en la Figura 4 (b) representa la población que ha caído en esa neurona. Con ello obtenemos una subdivisión del grupo Verde en cuatro subgrupos (V1, V2, V4 y V6) y del grupo Azul en tres (A3, A5 y A7).

La agrupación de las páginas por la red SOM está realizada en función a la similitud de las verificaciones.

Como es lógico un sitio web no tiene porque pertenecer en su totalidad a un solo grupo, puede ser que ciertas páginas del sitio Web tengan un comportamiento en relación a la accesibilidad y otras páginas otro.

El grupo R8, que es el que estaba en la zona roja, está formado por un porcentaje pequeño de páginas totales (Porcentaje de Agrupación 7%), y que sólo en pocos sitios Web pasa a ser un porcentaje superior al 20% de las páginas del sitio web. El grupo más numeroso es el A7 y V4 seguido de A5 y V6 y el resto tienen un porcentaje muy bajo de muestras.

5. Características de los grupos detectados

En este apartado se procede a buscar las verificaciones que caracterizan el comportamiento de cada uno de los grupos y subgrupos identificados por la SOM.

En la figura Figura 5 se muestra el mapa SOM con la U-matrix (matriz de distancias) y los componentes de cada una de las verificaciones (48) con las que se ha analizado.

Este mapa es el que ha permitido la identificación de los grupos y subgrupos comentados en el apartado anterior.

Sobre este mapa se procede a analizar los componentes que tienen un comportamiento similar o que definen el comportamiento del grupo etiquetado.

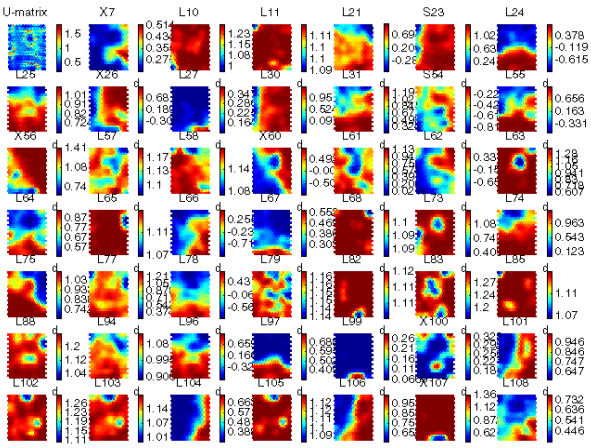


Figura 5. Mapa SOM con todos los casos de sitio Web

Se plantea la búsqueda de un conjunto de reglas que permitan definir el comportamiento del conjunto de datos aportado. Para ello, partiendo de los grupos detectados utilizando técnicas de agrupación, se pretende conseguir unas reglas que definan el comportamiento de los ocho grupos identificados.

Teniendo en cuenta que los grupos es una aproximación inicial al problema y que no tienen porque ser totalmente cerrados, se plantea realizar un árbol que defina el comportamiento genérico del conjunto de datos. Es decir, conseguir un conjunto de reglas, que pudiendo tener error de clasificación, definan el comportamiento general de los datos en vez del particular. Se ha de tener en cuenta que la frontera de los grupos se ha definido con técnicas de distancias entre los vecinos, con lo que al definir las reglas esas fronteras se pueden desplazar pudiendo tener paginas en las que se cometa un error de clasificación al asociar el grupo al que pertenece.

Si por el contrario hacemos un árbol que defina a la perfección los conjuntos de datos podemos estar añadiendo información errónea o excesivamente particular al conjunto de reglas.

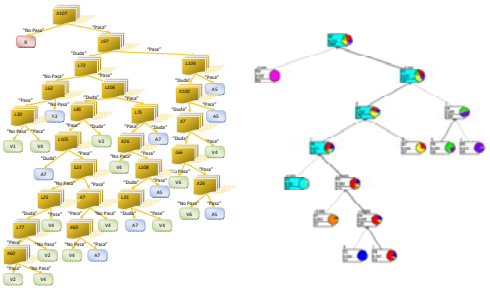


Figura 6. Reglas detalladas para la clasificación de los Grupos vs. Reglas generales

Si se realiza un conjunto de reglas como el que se ve en la figura anterior, el error cometido en la clasificación de los grupos es mínimo. Pero dado que los grupos no han sido dados en el conjunto inicial sino definidos mediante técnicas de agrupación, podemos estar incorporando al árbol el error generado en la definición de los grupos. Además, para cada uno de los grupos dispondríamos de varios caminos.

Por lo tanto se propone un árbol que pudiendo cometer un error de clasificación defina mejor el comportamiento de los datos con los que se trabaja, dado que es uno de los objetivos de este trabajo.

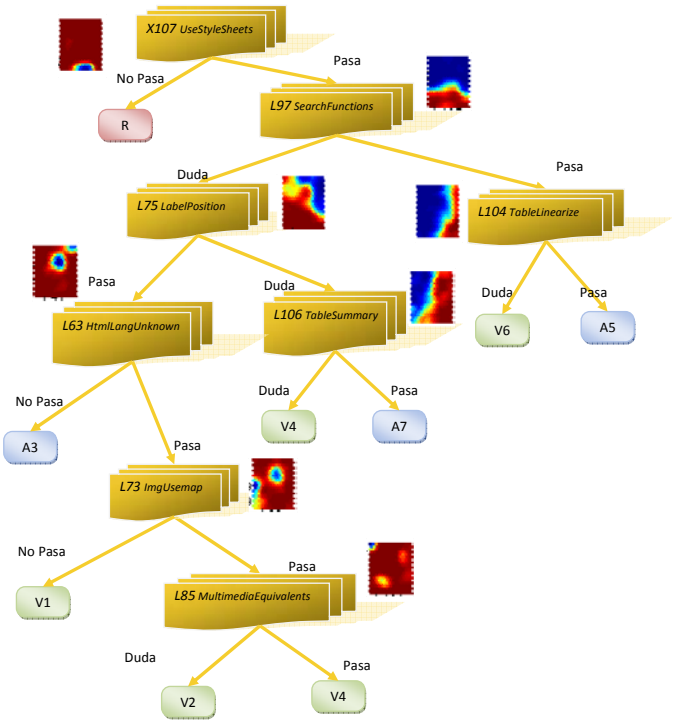


Figura 7. Reglas generales para la clasificación de los Grupos con la nueva verificación

El esquema que se muestra en Figura 7 representa el conjunto de reglas mínimo que permite definir los grupos detectados por la técnica de agrupación SOM.

Con este árbol se comete un grado de error superior al cometido por el detallado, pero éste nos permite definir las verificaciones que clasifican cada uno de los grupos encontrado. Esta clasificación comete ciertas discrepancias de agrupación que comentaremos más adelante.

Si tomamos como referencia la clasificación genérica realizada por el anterior conjunto de reglas reducidas, podemos superponer los resultados obtenidos por este, sobre el mapa SOM inicial.

Las líneas negras de la Figura 8 representan la división inicial realizada sobre los grupos detectados por el SOM y cada celda se ha marcado del color del grupo mayoritario detectado por las reglas generadas en la Figura 7.

Como se puede observar las zonas limítrofes entre grupos han cambiado debido a que se han utilizado pocas verificaciones para definirlo, lo que ha provocado que se generen errores identificados por:

- Celdas de un color no correspondiente a la zona.
- Zonas limítrofes que se prolongan o desplazan.

Como se puede observar sólo con 8 verificaciones somos capaces de reproducir las agrupaciones detectadas por el mapa SOM. Cabe destacar que el grupo A7 se prolonga y se inserta sobre V4, y en la parte central se ve una indecisión entre los grupos V6, V4, A5 y A7, debido a su gran similitud.

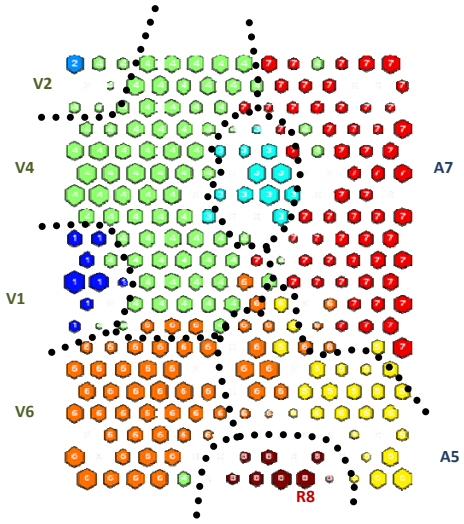


Figura 8. Calcificación realizada con el árbol nuevo vs. Calcificación de la SOM

Para concluir se muestra la representación del árbol en forma de condiciones:

R8	**Si**	X107 No pasa
A5	**Si**	X107 Pasa & L97 Pasa & L104 Pasa
V6	**Si**	X107 Pasa & L97 Pasa & L104 Duda
A7	**Si**	X107 Pasa & L97 es Duda & L75 es Duda & L06 Pasa
A3	**Si**	X107 Pasa & L97 es Duda & L75 Pasa & L63 No pasa
V1	**Si**	X107 Pasa & L97 es Duda & L75 Pasa & L63 Pasa & L73 No pasa
V2	**Si**	X107 Pasa & L97 es Duda & L75 Pasa & L63 Pasa & L73 Pasa & L85 es Duda
V4	**Si**	X107 Pasa & L97 es Duda & L75 Pasa & L63 Pasa & L73 Pasa & L85 Pasa
V4	**Si**	X107 Pasa & L97 es Duda & L75 es Duda & L06 es Duda

6. Conclusiones

Uno de los primeros puntos a destacar es que hay sitios Web que contienen un conjunto de páginas muy homogéneas en relación a las verificaciones de accesibilidad, lo que nos aporta información sobre la filosofía de desarrollo utilizada para realizar el portal. La presencia de homogeneidad en cuando a las verificaciones también pueden venir asociadas al uso de herramientas de desarrollo Web y no tanto a una filosofía de creación de la Web.

Por el contrario, también hay sitios Web con un comportamiento más heterogéneo, y este comportamiento puede estar asociado al propio mapa del sitio ya que puede contener información muy dispar.

Con los resultados obtenidos se puede observar que utilizando los métodos comentados en este trabajo se pueden identificar las verificaciones que definen el uso de las pautas de accesibilidad de los sitios Web.

Atendiendo a los grupos identificados dentro de este estudio se destaca la presencia de un grupo aislado del resto, como es el caso del etiquetado con el nombre Rojo. Como se puede observar en el árbol de clasificación de los grupos, éste se caracteriza por que la verificación “UseStyleSheets” no pasa, es decir, no se usa hoja de estilos.

En relación a los grupos Verde y Azul no está tan definida la diferencia, pero en el árbol se identifica que uno de los criterios para separar los dos grupos mayoritarios es la verificación

"TableSummary", la cual no aporta información sobre el contenido de la tabla. No pasar esta verificación puede ser debido a que muchos generadores de contenidos Web utilizan las tablas como organizador de la estructura de la página Web con lo que esa gran presencia de tablas no aporta contenido, ya que sólo tiene una misión estructural.

Una línea de futuro interesante a seguir sería aportar el conocimiento experto a la separación de la calidad en cuanto a la accesibilidad de un sitio Web, dado que algunas de las verificaciones generan como salida "duda". Debido a que por medio de los procedimientos automáticos no es posible conocer el cumplimiento de la pauta, se podría realizar por parte del experto una clasificación de un sitio o página web que cumple los criterios de accesibilidad, con lo que conseguiríamos transformarlo en otro problema de minería de datos que podría solucionarse con un modelo supervisado al disponer de un atributo definido.

Referencias

[1] WAI, *Web Accessibility Initiative* (2008), URL: http://www.w3.org/WAI/.

[2] Villanueva Balsera j., Rodriguez Montequín V., Alba Gonzalez-Fanjul C., Alonso Álvarez.C., "Estudio de Accesibilidad Web a través de técnicas de DataMining", *XI Congreso Internacional de Ingeniería de Proyectos Engineering*, 2007.

[3] *CRISP-DM 1996, CRoss-Industry Standard Process for Data Mining.*

URL: http://www.crisp-dm.org/.

[4] Friedman, J. H. (1991a). Multivariate Adaptive Regression Splines (with discussion), *Annals of Statistics*, 19, 1-141.

[5] Kohonen, T., 2006. Self-organizing neural projections. Neural Networks. 19 (6-7), 723-733.

[6] W. Chisholm, G. Vanderheiden, and I. Jacobs, eds, (1999), WCAG "Web Content Accessibility Guidelines 1.0". URL: http://www.w3.org/TR/WAI-WEBCONTENT.

Correspondencia (Para más información contacte con):

Silvia Barros Alonso
Universidad de Oviedo
Área de Proyectos de Ingeniería
c/ Independencia, 13
33004 Oviedo, Asturias (España).
Phone: +34 985 10 42 72
Fax: + 34 985 10 42 56
E-mail: secre@api.uniovi.es
URL: http://www.api.uniovi.es