UNIVERSIDAD NACIONAL DE LA PLATA

FACULTAD DE INFORMÁTICA

TESINA DE LICENCIATURA

Titulo: Análisis de Patrones en la Evolución de Wikis
Autores: Jonathan Martin
Director: Diego Torres
Codirector: Alejandro Fernández
Carrera: Licenciado en Sistemas

Resumen

Una wiki es un sitio web de edición colaborativa donde cualquier usuario puede editar, crear o modificar contenido directamente usando su navegador. En este trabajo se plantea un estudio sobre la evolución en la generación de conocimiento en el contexto de wikis. Al ser netamente colaborativas, reúnen la participación conjunta de muchas personas, en algunos casos miles o millones, que comparten un mismo objetivo. Es por ello y por ser una plataforma de fácil acceso, orientada a la generación de conocimiento y de gran actividad, que resultan un campo de gran interés. Además, la gran cantidad de personas que participan colaborativamente, por ejemplo, en las wikis nucleadas por Mediawiki, hacen que sea un fenómeno en el que se combinan personas con diferentes formaciones, ubicaciones geográficas y culturas. Por otro lado, las wikis poseen un “registro de revisiones” en el cual se guardan el momento y el autor de cada uno de los cambios realizados. En base a lo anterior se presenta un enfoque para poder analizar la evolución en el contenido una wiki a través de sus registros de revisiones. Este estudio nos permite detectar y clasificar patrones sobre la evolución en el contenido de la wiki. Este trabajo se centra en el estudio aplicado a la tecnología MediaWiki, motor de las wikis más importantes de la actualidad como Wikipedia.

Palabras Claves

Evolución en Wikis, Análisis Estructural, Patrones, Visualización, MediaWiki

Trabajos Realizados

Se desarrolló un prototipo que permite la extracción y visualización de información de los artículos de Wikipedia y sus revisiones. A través del mismo, se realizó un análisis de esas visualizaciones y se presentaron conclusiones sobre la relación de los datos obtenidos y los sucesos en el mundo real que los desencadenaron. Partiendo de estos análisis, se logró un nuevo enfoque para la detección de vandalismo en artículos de Wikipedia. Además, con la información generada por el prototipo se llevaron a cabo estudios con el entorno R para realizar análisis estadísticos de datos utilizando técnicas de aprendizaje no supervisado, de los cuales, se obtuvo una clasificación para las revisiones y patrones en la forma de desarrollar las actividades según su categoría.

Conclusiones

Se detectaron patrones dentro de las revisiones de los artículos de Wikipedia que permiten clasificar a los mismos en categorías, las cuales se caracterizaron con el fin de facilitar una posterior detección automatizada de las mismas. Además, se encontraron patrones respecto a la forma en que evoluciona la estructura de los artículos tanto en elementos como en las magnitudes en las que son aplicados. Por último, también se obtuvo un enfoque diferente para la detección de vandalismo en los artículos basándose en los cambios estructurales dados por el lenguaje de marcado utilizado en las MediaWikis.

Trabajos Futuros

Generar un sistema de detección automática y etiquetado de los tipos de revisiones hallados, para posteriormente integrarlo con estudios que analicen el comportamiento de los distintos tipos de usuarios durante su actividad y de la temática de la actividad. De esta integración obtener un sistema que permita recomendar información de interés para el tipo de actividad que se encuentre realizando el usuario en ese momento. También se espera que se puedan generar nuevas herramientas de detección automatizada de vandalismo a partir del nuevo enfoque desarrollado en este trabajo.

Fecha de la presentación: 02/2017

A mis padres y mi familia que me acompañaron a lo largo de todo este camino.

Agradecimientos

Quiero comenzar agradeciendo a mi director y a mi co-director de la tesina: Diego Torres y Alejandro Fernández. Quiero agradecerles a ambos por el apoyo, el entusiasmo y la guía que me brindaron hasta ahora y con la cual espero seguir contando en un futuro.

También quiero agradecerle a los chicos del LIFIA por el trato cálido y acogedor con el que me recibieron y acompañaron durante este tiempo.

Quiero agradecerle a la Facultad de Informática y a todos los profesores que tuve durante mi carrera, agradecerles por una educación de calidad y la dedicación con la que me enseñaron y continúan enseñando a muchos otros alumnos.

Ademas quiero agradecer a todos mis amigos y a mis compañeros a lo largo de la carrera, dado que sin ellos, la carrera no habría sido la maravillosa experiencia que fue.

Por ultimo y no menos importante quiero agradecer a mi familia que me apoyaron a lo largo de toda la carrera, que aprendieron con migo, se desvelaron acompañándome, me escucharon y siempre buscaron de ayudarme de todas las formas posibles. Muchas gracias.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Presentación

Las comunidades de construcción de conocimiento conocidas como wiki, por ejemplo Wikipedia(1) de la familia de MediaWiki(2), son sitios web de edición colaborativa donde cualquier usuario, sin estar necesariamente registrado, puede editar o crear nuevo contenido.

Este tipo de comunidades que permiten la generación dinámica de contenido poseen gran cantidad de aportes que consisten en cambios que se realizan sobre el contenido ya existente o creación de nuevo contenido accesible y modificable por la comunidad.

Estos cambios los almacenan en forma de revisiones en un artículo para poder recuperarse de cualquier cambio a un estado anterior y permitir rastrear de dónde proviene esta información.

En este capítulo hablaremos sobre el por qué estudiar la evolución de este tipo de comunidades de construcción de conocimiento y también sobre qué se espera obtener a partir de este estudio. Por último se describe cómo será la estructura a seguir del resto del trabajo.

1.2. Motivación

La construcción del conocimiento puede verse como resultado del proceso activo del aprendizaje(3). Siguiendo este concepto en la actualidad podemos encontrar varias comunidades que se forman a partir de grupos de personas con intereses comunes para compartir y generar conocimiento, como pueden ser foros especializados, o grupos como los brindados por el servicio de Groups Yahoo(4), o comunidades de la fami-

lia MediaWiki como Wikipedia. A estas comunidades se las denomina Comunidades de Construcción de Conocimiento (CCC), que son según Stahl(5) las que dan soporte computacional a las distintas etapas de la construcción de conocimiento.

En este trabajo se trata el caso particular de wikis como CCC, donde cualquier persona siendo o no un usuario registrado puede colaborar editando o creando artículos, categorías, páginas de discusión o de usuario, entre otros en la wiki. Esto genera la problemática de qué se debe saber si la información es de confianza o no, además como explican Viégas et. al.(6) en su trabajo la posibilidad de actos vandálicos injustificados contra el contenido de las wikis.

Para realizar un control de los cambios, las wikis proveen un registro de todas estas colaboraciones llamado "registro de revisiones". A partir de este registro un usuario puede observar quien realizó un cambio y cuando. Además la wiki provee una funcionalidad de seguimiento que permite a los editores interesados en un artículo estar al tanto de cuando se realiza un cambio. Entre ambas herramientas que brinda la wiki, y gracias a la comunidad, se logra mantener un control contra los actos vandálicos.

Por otro lado, el problema de la confianza de la información hace referencia al provenance o proveniencia de la información. Por proveniencia de la información nos referimos a las fuentes de la información, y a los procesos que influyen en la modificación o combinación de la misma. Para el estudio de la proveniencia de la información es de gran importancia determinar qué información es verdadera o confiable y cómo fue obtenida de una o más fuentes.

En este trabajo buscamos poder visualizar la evolución de los artículos de una wiki y poder a partir de la información obtenida detectar patrones o conductas repetidas entre los mismos.

1.3. Objetivo

Este trabajo tiene como objetivo visualizar la evolución de artículos de una wiki, a partir de lo cual detectar y clasificar patrones sobre la evolución en el contenido de la wiki. Se plantea analizar ciertos cambios estructurales como la creación de páginas o nuevos links, como así también la identificación de cambios e interpretaciones. Finalmente, generar diferentes visualizaciones que muestran la evolución a partir de gráficos e información estadística.

Los objetivos específicos de este proyecto son:

- Proponer una estrategia de extracción de información histórica de los cambios

realizados en una wiki de la familia MediaWiki.

- Plantear una estrategia de visualización de cambios entre dos revisiones.
- Exponer de forma gráfica desarrollo del contenido de la wiki en el tiempo.
- Brindar información gráfica y textual de cómo se organiza y distribuye el contenido de la wiki según aspectos propios de la misma.
- En base a la información analizada detectar y clasificar patrones sobre la evolución del contenido de la wiki.

1.4. Desarrollos propuestos

Se desarrollará un software que permita en primera instancia la obtención de la información histórica de una wiki a partir de los dumps, o copias de respaldo, en xml generados mensualmente.Con esa información se propondrá un modelo basado en el paradigma orientado a objetos que permita el desarrollo de estrategias para:

- Presentar información de cambios entre diferentes revisiones de artículos.
- Visualizar de forma gráfica y textual información sobre el estado al momento del dump de la wiki respecto a su contenido y su distribución según aspectos como autoría o estructurales como pertenencia a sus respectivos namespaces y categorías.
- Informar sobre la evolución del contenido de la wiki en el tiempo.

1.5. Contribuciones

En este trabajo se presenta un prototipo que permite la extracción y visualización de información de los artículos de wikipedias y sus revisiones. Con este prototipo se presentan análisis de esas visualizaciones que permiten observar un enfoque diferente para la detección de vandalismo en los artículos de las wikis, además nos permiten observar la relación de sucesos reales con la información obtenida.

También se presentan estudios realizados a partir de técnicas de aprendizaje no supervisado como cluster analysis, análisis de grupos en español, y minado de reglas de asociación. De los estudios realizados se presentan patrones dentro de las revisiones de los artículos de wikipedia que permiten clasificar a los mismos en categorías, las cuales se caracterizaron con el fin de facilitar una posterior detección automatizada de las mismas.

Además, se encontraron patrones respecto a la forma en que evoluciona la estructura de los artículos tanto en elementos como en las magnitudes en las que son aplicados.

1.6. Estructura de la tesis

A continuación se describe la estructura de este trabajo de tesis a partir del Capítulo 2 en el que se presentará el estado actual de las tecnologías relacionadas con esta tesina. Luego el trabajo se dividirá en dos partes, una de obtención de información y visualización y otra de análisis de los datos obtenidos.

La Primera Parte de obtención y visualización de la información comenzará en el Capítulo 3 donde se presenta la estrategia a seguir para la obtención de la información y las formas de visualización. El Capítulo 4 contendrá detalles de la implementación del prototipo que nos permite realizar esta tarea así como de las tecnologías que lo componen. En el Capítulo 5 se presenta información sobre las pruebas y resultados obtenidos a partir del prototipo.

La Segunda Parte de análisis de datos comenzará en el Capítulo 6 donde se explicará el proceso de análisis de datos a seguir. En el Capítulo 7 se presentarán los distintos análisis realizados con sus resultados.

Por último, en el Capítulo 8 se presentará una conclusión de la tesis y propuestas de trabajos futuros.

Capítulo 2

Marco teórico y estado del arte

2.1. Introducción

Las wikis son sitios web de edición colaborativa, la idea de los mismos es ser un compendio de conocimiento generado por la comunidad que la habita en principio sin ningún tipo de restricción.

> The problem with Wikipedia is that it only works in practice. In theory, it can never work.[El problema con Wikipedia es que solo funciona en la práctica. En la teoría, esta jamas funcionaría.] (7)

Esta cita hace referencia al grado de libertad que existe dentro de las wikis generando dificultades para controlar la comunidad o el contenido generado, por no decir que es teóricamente imposible por como están planteadas lograr un control total sobre la actividad en la misma. El hecho por lo tanto de que puedan existir y con tanto éxito es lo que ha generado una gran cantidad de estudios sobre las mismas entre las distintas comunidades de creación de conocimiento existentes.

En este capítulo se hablará primero sobre el concepto de wiki y las características o funcionalidades que brindan. Luego se hablará sobre diferentes estudios que ya se han realizado sobre las wikis y que serán usados como marco de esta tesina.

2.2. Wiki

Una wiki es un sitio web de edición colaborativa la cual permite a los usuarios crear y editar contenido de forma dinámica. La principal característica de las wikis es la flexibilidad con la cual se puede generar este contenido de forma simple a partir de un editor de texto que brinda la wiki que utiliza un sistema de markup o marcado de texto

para brindar formato al contenido o estructurarlo.

La primera wiki fue creada por el programador Ward Cunningham a la cual llamó WikiWikiWeb (8), su intención al crear este software fue permitir mejorar la generación de documentación de software a partir de una edición sencilla desde un editor de texto que permitiera la creación y modificación de contenido de forma colaborativa y rápida.

Por otro lado con los avances de la tecnología a finales del siglo XX la idea de enciclopedia la cual ya había evolucionado de las típicas enciclopedias en libros a enciclopedias contenidas en software como el Encarta de Microsoft tuvo una nueva reforma y comenzó a gestarse la idea de enciclopedias en Internet.

El primer proyecto siguiendo esta idea fue el de Nupedia (9), fundado por Jimmy Wales y con Larry Sanger como editor en jefe, en el cual se juntó un amplio grupo de especialistas para escribir el contenido de la misma. Una de las principales diferencias con las wikis de hoy en día era que Nupedia tenía 7 niveles de comprobación para la aprobación de una edición lo cual generó que la creación de contenido fuera extremada lenta llegando a producir solo 12 artículos durante su primer año.

A partir de charlas entre Wales y Sanger sobre cómo generar contenido de forma más dinámica se originó la idea de aplicar tecnología wiki con dicho fin. Así aunque la comunidad de Nupedia se mostraba bastante resistente al cambio se creó una de las wikis más conocidas en la actualidad y sobre la cual se trabajara en este trabajo: Wikipedia.

Wikipedia al igual que cualquier wiki que utiliza el motor MediaWiki permite a los usuarios crear contenido nuevo de forma rápida desde un editor de texto que las mismas proveen, luego dicho contenido es inmediatamente publicado con el título de la página como identificador por lo cual no pueden existir dos páginas o artículos,dentro de una wiki son sinónimo, con el mismo título.

Esta posibilidad de editar libremente, aunque en la actualidad las páginas que puedan tener un impacto social importante suelen estar protegidas, puede generar que algún usuario edite de forma errónea el contenido de un artículo por lo cual los demás usuarios deberán corregir el contenido del mismo. Además hoy en día las wikis posee bots o sistemas automatizados para detectar algunos tipos de ediciones erróneas y revertirlos de forma automática.

En la Figura 2.1 podemos observar el artículo de Nupidia en la Wikipedia en inglés, dentro del mismo vemos que el usuario tiene la posibilidad de editar el contenido seleccionando la opción “edit”. Dicha funcionalidad nos brinda un editor de texto como el que se muestra n la Figura 2.2 donde podemos observar el contenido del artículo y modificarlo.

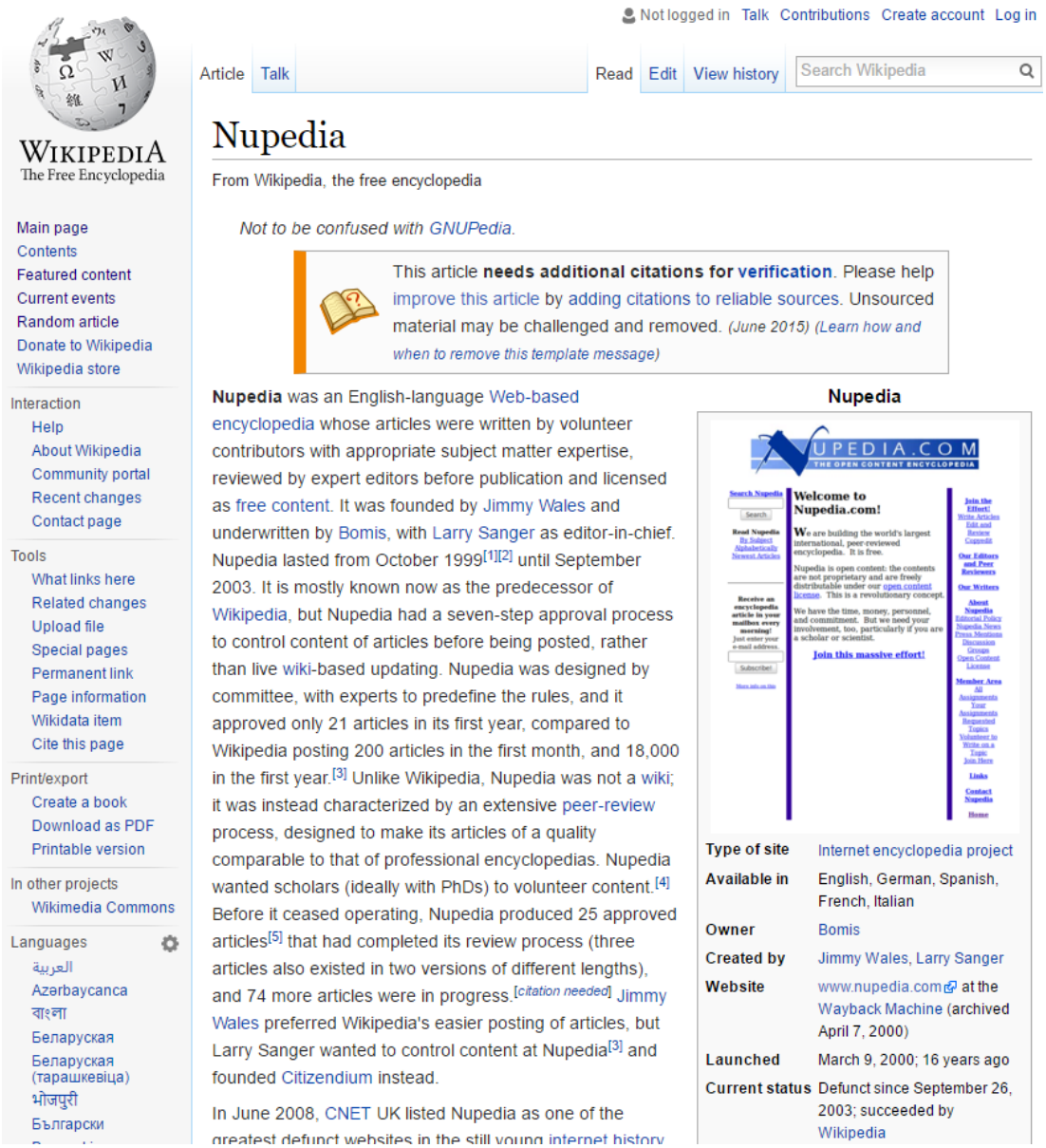


Figura 2.1: Artículo de Nupedia de en.wikipedia.org.

En la Figura 2.2 podemos observar además del editor el lenguaje de markup brindado por las wikis para dar formato a los artículos donde podemos ver por ejemplo enlaces internos a otros artículos en la wiki definidos entre doble corchetes [[ ]] conteniendo el título del artículo como por ejemplo un enlace interno al fundador de Nupedia de la siguiente forma [[Jimmy Wales]].Otros elementos de markup que se utilizaran en este trabajo son los de la Tabla 2.1 donde podemos observar su nombre, sus símbolos de apertura y cierre aplicados en un ejemplo y una descripción. Al igual del ejemplo brindado con el enlace interno al nombre del fundador de Nupedia, cada vez que se utiliza un markup sus efectos se aplican al texto contenido entre su apertura y su cierre o en caso de no tener un cierre hasta el próximo salto de linea.

Luego de que un usuario finaliza su edición y la guarda la misma es automáticamente visible en el artículo como la versión final hasta que se realice otra edición.

Dado que se pueden generar casos en los que se generen ediciones erróneas o las consideradas ediciones vandálicas, que es toda edición con intención maliciosa, las wikis poseen por cada artículo un historial de todos los cambios que se realizaron.

En la Figura 2.3 podemos ver este historial de cambios conocido como “Historial de Revisiones”. El usuario puede acceder a dicho historias a partir de la opción “View History”, dentro puede observar un listado con las revisiones que se realizaron de la más reciente a la más antigua. Esta historia posee información sobre cuándo se realizó, quien la realizó y el contenido final de la misma. También nos permite seleccionar dos revisiones para comparar el contenido de ambas a fin de visualizar los cambios entre dos revisiones.

Por último en las wikis para organizar el contenido y facilitar la navegación se utiliza un tipo de artículo especial llamado categoría, una categoría permite agrupar un conjunto de artículos relacionados por el significado de la categoría. Por ejemplo el artículo de Nupedia podemos ver en la Figura 2.4 que se encuentra en la categoría “Free internet encyclopedias”.

En esa categoría podemos encontrar otros artículos sobre enciclopedias libres en Internet como puede ser el artículo de Wikipedia. Además las categorías pueden contener subcategorías para mejorar el orden y agrupación de los artículos. Dado que las categorías son editadas por los mismos usuarios puede suceder que se produzcan bucles o errores en la estructura de las categorías.

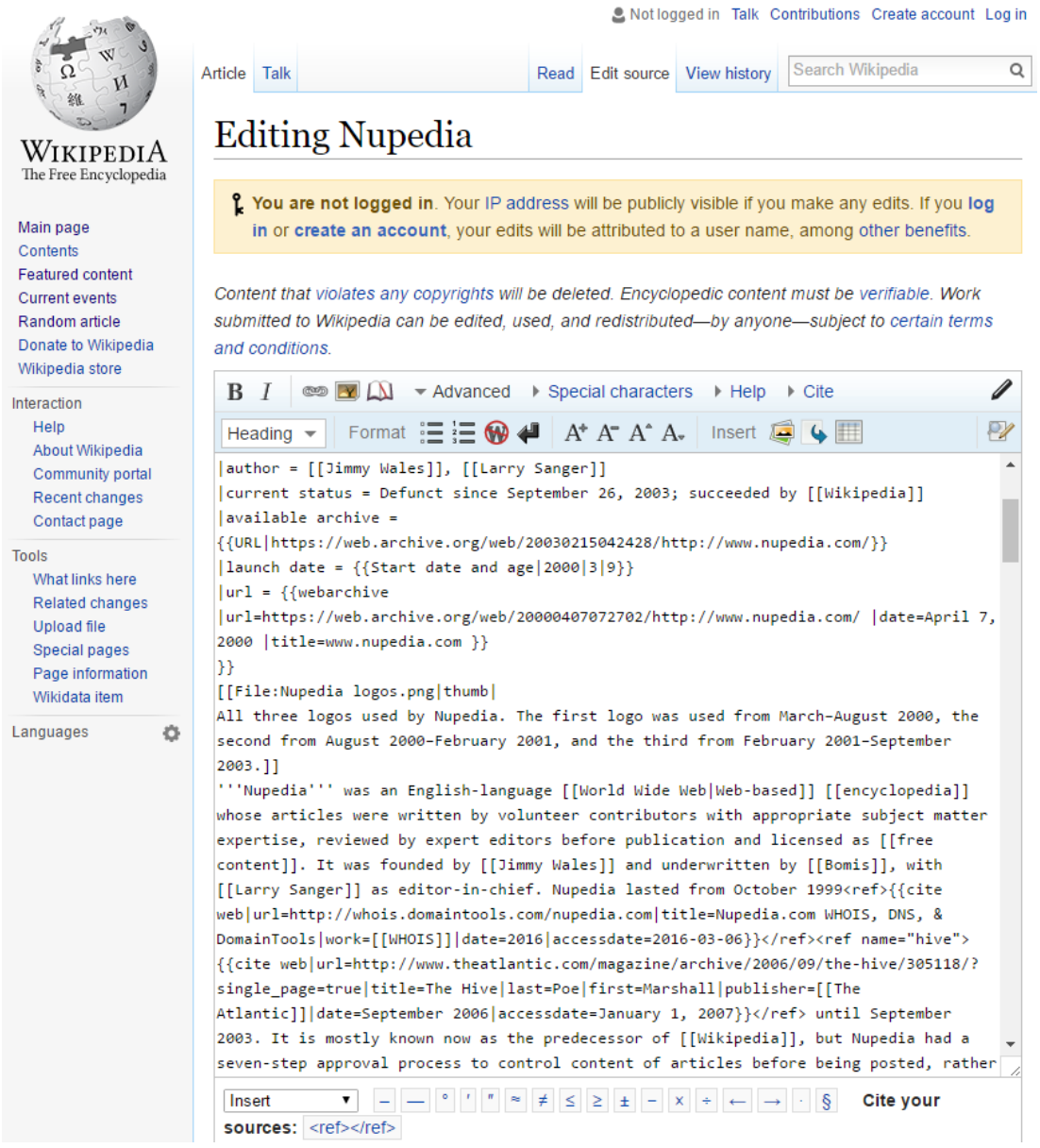


Figura 2.2: Editor del artículo de Nupedia de en.wikipedia.org.

Nombre	Apertura	Cierre	Descripción
Nowiki	<nowiki>	</nowiki>	Bloquea la aplicación de formatos en su interior
Grande	<big>	</big>	Aumenta el tamaño del texto
Pequeño	<small>	</small>	Disminuye el tamaño del texto
Superíndice	[	]	Crea un superíndice
Subíndice	_		Crea un subíndice
Tachado	<s>	</s>	Tacha el texto
Bloque de Cita	<blockquote>	</blockquote>	Genera una cita enmarcada
Includeonly	<includeonly>	</includeonly>	Se utiliza para el mantenimiento de templates
Referencia	<ref>	</ref>	Crea una referencia
Encabezado de 2.º nivel	Dos iguales ==	Dos iguales ==	Genera un título tipo 2
Encabezado de 3.º nivel	Tres iguales ===	Tres iguales ===	Genera un título tipo 3
Encabezado de 4.º nivel	Cuatro iguales ====	Cuatro iguales ====	Genera un título tipo 4
Encabezado de 5.º nivel	Cinco iguales =====	Cinco iguales =====	Genera un título tipo 5
Cursiva	Dos comillas simples ‘’	Dos comillas simples ‘’	Aplica al texto cursiva
Negrita	Tres comillas simples ‘’’	Tres comillas simples ‘’’	Aplica al texto negrita
Negrita & cursiva	Cinco comillas simples ‘’’’’	Cinco comillas simples ‘’’’’	Aplica al texto cursiva y negrita
Enlace Externo	Corchete simple [	Corchete simple]	Genera un enlace externo a la wiki
Enlace Interno	Corchete doble [[	Corchete doble]]	Genera un enlace a contenido de la wiki
Elemento Enumerado	Numeral al comienzo de cada elemento #	No posee	Permite crear listas numeradas
Elemento Listado	Asterisco al comienzo de cada elemento *	No posee	Permite crear un listado punteado
Redireccion	#REDIRECT [[	Corchete doble]]	Genera una redirección a otro artículo
Sangria 2	Doble dos puntos ::	No posee	Genera sangría de tipo 2
Sangría 1	Dos puntos :	No posee	Genera sangría de tipo 1
InfoBox	{{Infobox	Doble llave }}	Genera una estructura de información conocida como Infobox
WikiTable	{— class=ẅikitable	Barra y llave —}	Crea una tabla
Cita	{{cite	Doble llave }}	Crea una cita

Tabla 2.1: Elementos de markup

Figura 2.3: Historial de revisiones del artículo de Nupedia de en.wikipedia.org.

Figura 2.4: Categorías del artículo de Nupedia de en.wikipedia.org.

2.3. Visualización de datos

La visualización de datos se utiliza para mejorar la forma en que se transmite la información compleja obtenida de estudios complejos a los usuarios, para facilitar la comprensión de la información resultante de análisis estadístico o también simplemente para facilitar la comprensión de los conjuntos de datos que se poseen.

Por lo cual se puede decir que visualización de datos es la disciplina que tiene como objetivo la comunicación clara y eficiente de información a los usuarios.(10)

Para cumplir con esta tarea se utilizan técnicas que permiten presentar la información de forma gráfica como gráficos de barra, de torta o dona, series de tiempo o cualquier representación gráfica que cumpla con la consigna de permitir comunicación clara y eficiente.

Dada la variedad de herramientas para realizar visualización existentes nos centramos en las que nos permitían implementar técnicas de visualización en interfaz web. Una de las herramientas estudiadas fue D3JS (11) que es un framework implementado como una librería javascript que permite introducir visualizaciones a nuestra aplicación y nos brinda decenas de tipos de visualizaciones. Otra herramienta que se estudio para su uso fue Chart.js (12) la cual aun que permitía menos tipos de técnicas de visualización que D3JS, brinda la ventaja de trabajar con HTML5 Canvas aumentando su rendimiento y compatibilidad con navegadores mas actuales. Por ultimo la herramienta de visualización que se decidió utilizar en este trabajo fue Google Charts (13), los motivos por lo que fue seleccionada son que:

- Permite implementar múltiples técnicas de visualización, mas que las que permite Chart.js.
- Brinda integración con HTML5.
- Nos brindan herramientas exploratorias de las visualizaciones que facilitan su estudio y análisis.

En este trabajo se utilizan varias de estas técnicas para permitirnos obtener una mejor visión de cómo es la evolución de los artículos en las wikis.

2.4. Proveniencia de la información

La proveniencia de la información según Hartig (14) representa sobre la información su historia desde su creación incluyendo información sobre sus orígenes. La proveniencia de la información por lo tanto se basa en la capacidad de rastrear las fuentes de la

información para poder medir la confiabilidad y utilidad de la información para el fin deseado.

La proveniencia de la información es utilizada para múltiples fines, desde aplicaciones de negocio al ámbito científico. Uno de los ámbitos en la que se utiliza proveniencia de la información es en el contexto de la web semántica la cual utiliza estas técnicas para brindar acceso a información de proveniencia tanto a sistemas computacionales como humanos.

Dado que en las wikis se genera una gran cantidad de información es de gran importancia la aplicación de estas técnicas como sucede en las wikis semánticas, las cuales son wikis que implementan tecnologías de la web semántica para hacer accesible su información tanto a sistemas como a personas.

En este trabajo se utilizaran los historiales de revisiones de cada articulo ,que brindan las wikis con información sobre que cambios se realizaron, quien los realizo y cuando, como fuente de información de proveniencia de la información contenida en el articulo.

2.5. Minado de patrones

Según Zumel et al. (15) uno de los usos comunes del minado de datos es la detección de patrones dentro de conjuntos de datos. Particularmente se buscan patrones que no son evidentes, para este fin se usan diversas herramientas de estadística y aprendizaje automático.

Según Saxena et. al (16) las principales técnicas para esta tarea son:

- Reglas de Asociación.
- Clustering también conocido como Cluster Analysis[Análisis de grupo].
- Caracterización y comparación
- Análisis secuencial de patrones.
- Análisis de tendencia.

Las técnicas que se utilizan en este trabajo son análisis de grupo para observar cómo la información se agrupa de forma automática, buscando los posibles patrones que muestran esas agrupaciones, o por ejemplo el minado de reglas de asociación que evalúan la co ocurrencia de los elementos de un conjunto de datos.

Este tipo de técnicas junto con las técnicas de visualización mencionadas con anterioridad son herramientas de gran importancia para todos los procesos de la ciencia de

Rol	Actividades centrales Característico
Social Networker	Participar en páginas de discusión, páginas de usuario y modificación de referencias.
Fact Checker	Eliminar información, eliminación de enlaces internos, eliminación de referencias, eliminación de archivos, eliminación de markup, eliminación de enlaces externos.
Substantive Expert	Agregar información, agregado de enlaces internos, agregado de referencias, agregado de archivos, agregado de markup, agregado de enlaces externos.
Copy Editor	Corregir gramática, corrección de frases, y relocacion de contenido.
Wiki Gnomes	Modificar de enlaces internos, inserción de templates, modificación del markup y discusión en las páginas de discusión de Wikipedia.
Valdal Fighter	Revertir revisiones, discusión en páginas de usuario, agregado de referencias y eliminación de enlaces externos.
Fact Updater	Modificar templates y modificaciones de referencias
Wikipedian	Agregar enlaces internos, modificaciones a los templates y inserción de archivos.

Tabla 2.2: Roles de usuarios

datos.

2.6. Trabajos relacionados

Dada las posibilidades que nos brindan las wikis como comunidades de creación de conocimiento es comprensible que haya un gran interés por el estudio de las mismas esto se ve reflejado en el gran número de estudios relacionadas a las mismas en diferentes aspectos.

Por ejemplo podemos encontrar estudios como el de Yang et al. (17) que estudian en Wikipedia como se pueden detectar los tipos o roles de usuarios según su actividad presentes en la Tabla 2.2

También explica cómo la participación en distintas medidas de cada tipo de editor

son necesarias para poder mejorar la calidad de los artículos y como la confianza que se le tiene a un autor se transfiere a los datos que aporta siendo estos por lo tanto información de confianza dentro de la comunidad.

En conjunto con la relación entre el autor y la calidad de sus aporte en Orlandi et al. (18) los autores describen como la obtención de la proveniencia de la información ayuda a asegurar la confianza de sus datos.

Otro estudio relacionado con la actividad de los editores se da en Geiger et al.(19) donde en lugar de contar los aportes de un editor como el número de ediciones que realiza explican como poder plantear la actividad del editor como una estimación de sus sesiones de edición y por lo tanto en lugar de tener el número de revisiones realizadas se posee el número de horas trabajadas por el editor.

Un trabajo que se relaciona tanto con las actividades de los autores como la visualización de la misma es la que encontramos en Viégas et al.(6) donde se describe cómo los autores publican y editan contenido colaborando en conjunto para consensuar las diferencias, en este trabajo utilizan la herramienta HistoryFlow la cual permite ver la actividad de los autores en los artículos. Con la información de la actividades de los autores la herramienta utiliza técnicas de visualización de datos, más exactamente histogramas, a fin de facilitar el entendimiento de lo que sucede en el artículo. Además en el trabajo utiliza la herramienta para de forma manual a partir de los gráficos poder detectar patrones. En este trabajo se encontraron patrones relacionados con la lucha contra el vandalismo que se refleja en los gráficos y también la negociación entre autores junto con la estabilidad del contenido.

En este trabajo también se espera que a partir del análisis del provenance se puedan hallar patrones o escenarios recurrentes y clasificarlos, sobre este tipo de análisis podemos ver en el trabajo de Duarte et. al.(2) un estudio en el cual se evalúa la participación de los editores en la wiki denominada WikiHaskell la cual es una wiki en la cual los alumnos crean material complementarios sobre bibliotecas del lenguaje de programación Haskell. En este estudio se plantea que a los editores se los puede clasificar según la forma de realizar sus aportes tanto en cantidad como en el momento en que comenzaron a hacerlo y por cuánto tiempo. Donde la clasificación es la siguiente:

- Perfil continuo: Considerado el perfil óptimo. El alumno va haciendo aportes de forma continua durante todo el desarrollo del trabajo.
- Perfil en escalón: Este es también un perfil bueno, el alumno va haciendo aportes de forma continua aunque algo intermitente.
- Perfil pico al principio: Este es el perfil del abandono, ya que lo siguen alumnos que sólo realizaron aportes al principio pero que después abandonaron el trabajo y la asignatura.

- Perfil incremento a mitad del periodo: Junto con el perfil en escalón, éste es el que más han seguido. En éste, la mayor parte del trabajo la realizan a mitad del periodo de desarrollo.
- Perfil pico al final: Este es el perfil del alumno que deja el trabajo para última hora.

Es importante destacar que las actividades de la wiki eran durante el desarrollo de un curso por lo cual se poseía un espacio de tiempo delimitado y se conoce exactamente quienes debían ser los autores, en este trabajo se desea trabajar sobre wikis más generales sobre las cuales pueda visualizarse su evolución desde su creación hasta la actualidad. Además en este trabajo se buscará encontrar patrones sobre la evolución del contenido en la wiki y su estructura y no solo en lo referente a los autores.

Parte I

Obtención y visualización de información

Introducción

En esta Parte del trabajo se habla de como se puede obtener la información contenida en las wikis, cuales son las fuentes de información y como se estructura la misma. También se hablara sobre las métricas que se utilizaran a lo largo del trabajo para medir la información.

Luego se presenta un primer prototipo que se desarrollo para la obtención automática de la información y la visualización de los valores resultantes de medir la información con las métricas. Se explicara que visualizaciones brinda este prototipo y con que métricas se relacionan cada uno.

Por ultimo se presenta un conjunto de datos para utilizar en el prototipo, y a la información y gráficas resultantes se las presenta y analiza. De este análisis se presentan algunos de los casos de interés mas relevantes que nos permiten observar como se reflejan los sucesos del mundo real en los gráficos y también se obtiene un nuevo enfoque para la detección de actos de vandalismo en artículos de Wikipedia.

Capítulo 3

Diseño del experimento

3.1. Introducción

Una wiki, implementada con la tecnología de MediaWiki, es un sitio web donde cualquier persona puede editar o crear contenido contribuyendo al crecimiento de la misma. Dado que cualquier persona puede editar, se conserva un historial de cada edición realizada a una pagina llamadas revisiones. Las revisiones conservan datos sobre quien la realizó, en qué momento y el contenido final de la página para esa revisión.

Con estas revisiones lo que se desea evaluar es cómo es la evolución de las páginas de una wiki y se desea poder visualizar dicha información para:

- Saber cuál fue el crecimiento o decrecimiento en cada revisión.
- Conocer qué autores y en qué medida contribuyeron a la página.
- Medir el nivel de actividad en la edición de una página en días.

Además las wikis proveen a los editores un lenguaje de marcado el cual permite darle formato y estructurar el texto de las revisiones. En este trabajo se propone utilizar este lenguaje para obtener información de la estructura del contenido de la revisión, a los cuales denominaremos estilos, y a partir de esto se espera:

- Poder analizar el cambio de estilos entre dos revisiones.
- Obtener información de los cambios de los estilos para una página a lo largo de su historial de revisiones, tanto de forma general como específica por cada tipo de estilo.

Luego a partir de la visualización de toda la información obtenida en este punto se desea poder detectar hitos o sucesos relevantes en las historias de las páginas y comprobar su relación con el contexto histórico durante el cual sucedieron.

3.2. Definiciones preliminares y métricas

A partir de la información obtenida se desea evaluar las mismas con un conjunto de métricas las cuales se utilizaran para la presentación gráfica de la información posteriormente.

Comenzaremos definiendo una wiki **W**,perteneciente a la familia de las MediaWikis, donde una wiki contiene un conjunto de páginas y namespaces tal que $\mathbf{W} = (\mathbf{N}, \mathbf{P})$ siendo **N** es el conjunto de namespace y **P** el conjunto de páginas.

Cada página de **P** contiene información del namespace al que pertenece, el cual nunca cambia una vez creada, y posee una colección de revisiones que le fueron realizadas de forma tal que definimos una Página **p**, donde $\mathbf{p} = (\mathbf{n}, \mathbf{R})$ con $\mathbf{p} \in \mathbf{P}$ y siendo **R** el conjunto de revisiones de la página y n el namespace de la página con $\mathbf{n} \in \mathbf{N}$.

Cada revisión de **R** contiene quien la realizó y cuando, el texto final de la revisión y si existía una revisión anterior también contiene referencia a la misma. Por lo cual podemos definir a la revisión $\mathbf{r} \in \mathbf{R}$ como $\mathbf{r} = (\mathbf{id}, \mathbf{pr}, \mathbf{d}, \mathbf{a}, \mathbf{t})$ con **id** como identificador que es un numero natural mayor a 0 y único para su identificación, y siendo **pr** la revisión anterior ,o padre, identificada por un numero natural mayor a 0 o negativo si **r** no posee revisión anterior , además siendo **d** la fecha en que se realizó la revisión, **a** el autor de la misma y **t** el texto de la página para dicha revisión.

Para referirnos al conjunto de revisiones de la página **p** utilizaremos $\mathbf{R_p}$ y para referirnos a una revisión particular de la página **p** utilizaremos $\mathbf{r_p}$ siempre que $\mathbf{r_p} \in \mathbf{R_p}$. También nos referiremos para una revisión **r** a su revisión anterior, autor, fecha en que se realizó y texto como $\mathbf{pr_r}, \mathbf{a_r}, \mathbf{d_r}$ y $\mathbf{t_r}$ respectivamente.

Ademas en este trabajo para referirnos al cardinal de un conjunto **B** se utilizara la notación $\mathbf{card}(\mathbf{B})$. A partir de esto se definieron las siguientes métricas:

#RevisionsOfAuthor(a,R):

Esta métrica indica la cantidad de revisiones realizadas por el autor a en la colección de revisiones **R**.

$$\mathbf{card}(\{\mathbf{r}/\mathbf{r} \in \mathbf{R} \wedge \mathbf{a_r} = \mathbf{a}\})$$

Con esta métrica podemos por ejemplo de un conjunto revisiones realizadas por los autores $\mathbf{a_1}$,$\mathbf{a_2}$,$\mathbf{a_3}$ saber cuantas revisiones realizo exactamente cada uno de los autores.

PercentOfAuthorRevisions(a,R):

Esta métrica representa el porcentaje de revisiones realizadas por el autor a de la colección **R** de revisiones.

$$\frac{\#RevisionsOfAuthor(a, R) * 100}{R}$$

Con esta métrica dado un conjunto de revisiones podemos saber del total cual fue el porcentaje de revisiones realizadas por determinado autor.

PercentOfRevisionByAuthor(p):

A partir de las métricas mencionadas anteriormente se puede calcular cuál fue el aporte de cada autor **a** una página **p** con la siguiente métrica.

$$\{(a_r, PercentOfAuthorRevisions(a_r, R_p))/r \in R_p\}$$

Con esta métrica para un determinado articulo podemos conocer por cada autor que realizo ediciones cual es el porcentaje de ediciones que hizo sobre el total que posee el articulo.

RevisionsOfDay(d,R):

Con esto podemos obtener las revisiones pertenecientes a la colección **R** tales que las mismas hayan sido realizadas en la fecha **d**.

$$\{r/r \in R \wedge d_r = d\}$$

RevisionPerDay(p):

Esta métrica nos permite obtener las revisiones que se realizaron para cada dia en el que al menos se realizó una revisión, además a partir de esto podemos saber para un día determinado la cantidad de revisiones que se realizaron.

$$\{(d_r, RevisionsOfDay(d_r, R_p)/r \in R_p)\}$$

```
== Uso de la clasificación TAS ==

Debe observarse, como se discute con detalle por Le Maitre y otros (2002), que esta clasificación
no puede aplicarse a todas las [[roca volcánica|rocas volcánicas]]. Ciertas rocas no pueden
nombrarse usando el [[diagrama]]. Para otras, se deben usar criterios adicionales mineralógicos,
químicos, y de textura, como por ejemplo con los [[lamprófiro]]s.
```

Figura 3.1: Extracto del articulo "Clasificación TAS"de Wikipedia.

BytesOfRevisionsPerDay(p):

Asumiendo que #**t** nos da el tamaño del texto en bytes esto permite obtener por día los cambios de longitud en el texto de una página **p**.

$$\{(d_r, \#t_r)/r \in R_p\}$$

Además de los componentes propios de las páginas que obtenemos de las wiki como las revisiones o namespaces, también extraemos de los textos de las revisiones una colección de estilos , nombrada **S**, tal que $\mathbf{S_t}$ está formada por un conjunto de tuplas que contienen un estilo **s** y las ocurrencias de dicho estilo en **t**.

$$S_t = \{(s, T)/s \in S \wedge T = \{t_i/t_i \in t\}\}$$

Por ejemplo a partir del fragmento de texto como el de la Figura 3.1. Podemos decir que el conjunto $\mathbf{S_t}$ contendría solo dos elementos uno representando a los Encabezados de nivel 2 y el segundo a los enlaces internos, y cada uno con su respectivo conjunto de ocurrencias.

#OcurrencesOfStyle(s,t):

Esta métrica permite obtener la cantidad de ocurrencias del estilo **s** en el texto **t**.

$$card(\{i_T/i_S \in S_t \wedge i_s = s\})$$

De esta forma podemos obtener métricas basadas en esta.

StylesInText(t):

Con esta métrica obtenemos un listado con la cantidad de ocurrencias de cada estilo en el texto **t**.

$$\{(s, \#OcurrencesOfStyle(s, t))/s \in S_t\}$$

StylesOfRevisionsPerDay(p):

Siendo $\#\mathbf{S_t}$ la cantidad de elementos de la colección $\mathbf{S_t}$ obtenemos por revisión la cantidad de estilos que posee y la fecha de la revisión.

$$\{(d_r, \sum_{i}^{StylesInText(t_r)} i_T)/r \in R_p\}$$

Capítulo 4

Desarrollando un prototipo

4.1. Introducción

Surgiendo de las métricas anteriormente definidas y de los medios de obtención de información que se definen en este Capitulo se dio comienzo a un prototipo utilizando el lenguaje de programación java con librerías para el desarrollo de aplicaciones web. Se optó por una implementación con interfaz web que brinda flexibilidad al momento de elegir qué herramientas de visualización utilizar, por lo cual se aliena con la primera intención de este prototipo que es dar visibilidad gráfica a la información y las métricas generadas.

Este prototipo en java utiliza las tecnologías de Spring MVC y Hibernate los cuales son descritos por Luna (20). También se utilizo la librería XStream (21) para facilitar el procesamiento de archivos xml. El prototipo nos permite:

- Cargar de dumps completos de MediaWikis a partir de archivos xml.
- Descargar y almacenar páginas individuales de la versión inglesa de Wikipedia.
- Obtener de las categorías de las páginas descargadas de forma individual.
- Obtener estadísticas generales de la wiki al ser cargada a partir de un dump completo.
- Obtener un listado de las páginas de las páginas almacenadas.
- Por cada página obtener un listado de revisiones y poder ver las diferencias entre las mismas.
- Por cada página obtener de forma gráfica información sobre todas sus revisiones.
- Por cada página obtener de forma gráfica de los cambios de estilos a lo largo del tiempo.

- Exportar información de los artículos en forma de json individuales.

A continuación detallan cuales son las posibles fuentes de información y los métodos que nos permitirán obtenerla. Luego se explicará con más detalle cada una de las tecnologías utilizadas y la motivación para el uso de las mismas. También se detalla sobre algunas de las funcionalidades listadas anteriormente en aspectos como implementación o funcionalidad.

4.2. Fuentes de información

Dado que toda la información que es de nuestro interés para analizar se encuentra dentro de las revisiones nuestro primer enfoque es cómo obtener las mismas para nuestro análisis.

En este trabajo se encontraron dos fuentes de información que nos brindaban información sobre el total de las revisiones para artículos tanto de forma individual utilizando la API Special:Export que brindan las wikis y para obtener todas las revisiones de todos los artículos de una wiki podemos optar por el procesamiento de dumps de la wiki.

A continuación se explicaran como funcionan ambas estrategias y sus ventajas y desventajas.

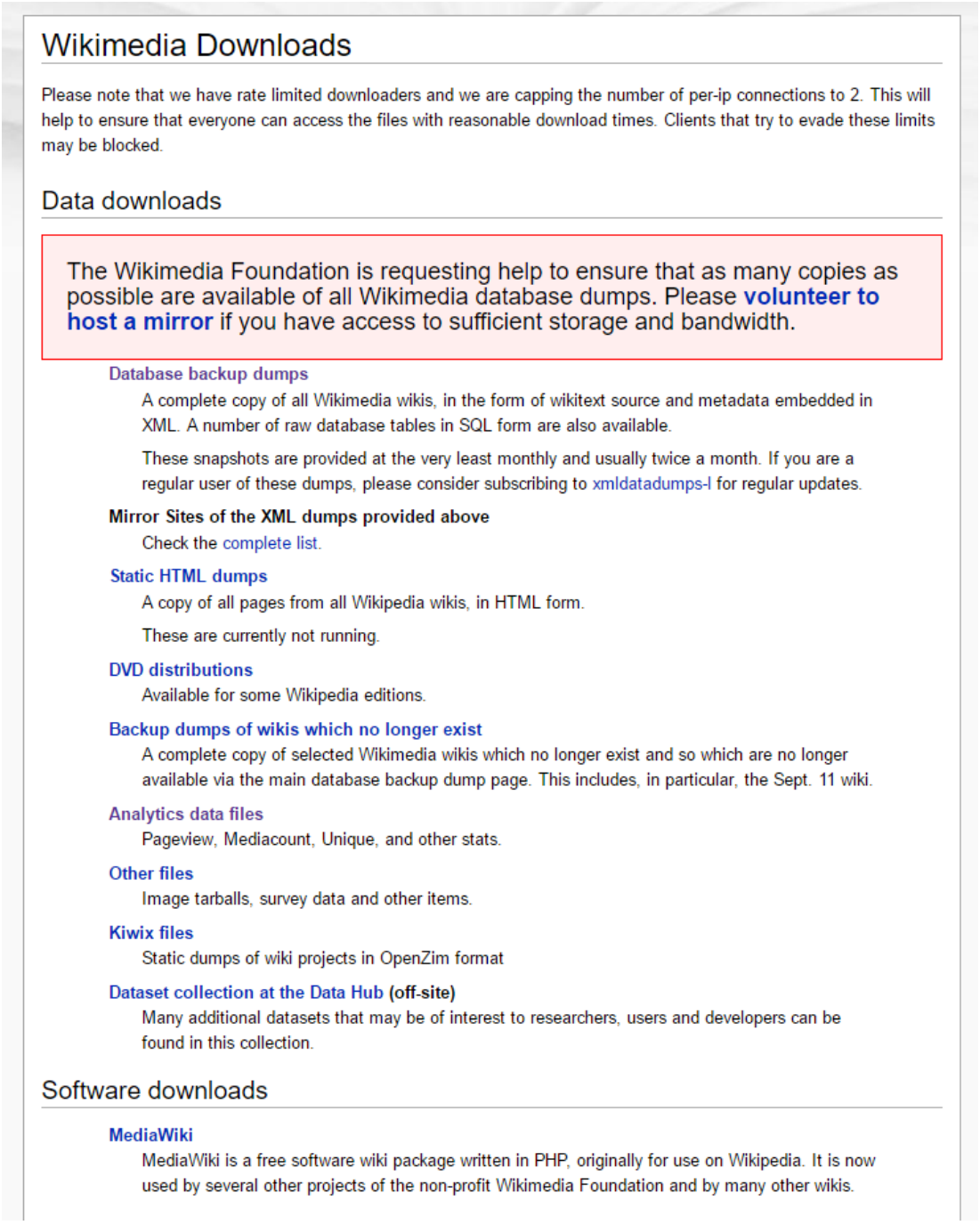
4.3. Respaldos de MediaWikis

El motor de wikis MediaWiki brinda la posibilidad de a las wikis de periódicamente exportar todo su contenido a forma de respaldo. Estos respaldos conocidos como dumps quedan accesibles desde la web de wikimedia dumps.wikimedia.org como observamos en la Figura 4.1 .

Desde esta web podemos podemos acceder a los respaldos de todas las MediaWikis accediendo a “Database backup dumps”, a partir de allí podemos seleccionar que wiki es de nuestro interés para obtener sus dumps.

Una vez seleccionada la wiki nos encontraremos con que no hay un solo tipo de dump sino que tenemos una gran variedad de archivos para descargar a continuación se describen algunos de los mismos.

Como se puede apreciar tenemos una variedad de archivos con gran variedad de información y distintos niveles de completitud. En nuestro caso nos interesa estudiar



Wikimedia Downloads

Please note that we have rate limited downloaders and we are capping the number of per-ip connections to 2. This will help to ensure that everyone can access the files with reasonable download times. Clients that try to evade these limits may be blocked.

Data downloads

The Wikimedia Foundation is requesting help to ensure that as many copies as possible are available of all Wikimedia database dumps. Please **volunteer to host a mirror** if you have access to sufficient storage and bandwidth.

Database backup dumps

A complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML. A number of raw database tables in SQL form are also available.

These snapshots are provided at the very least monthly and usually twice a month. If you are a regular user of these dumps, please consider subscribing to xmldatadumps-l for regular updates.

Mirror Sites of the XML dumps provided above

Check the complete list.

Static HTML dumps

A copy of all pages from all Wikipedia wikis, in HTML form.

These are currently not running.

DVD distributions

Available for some Wikipedia editions.

Backup dumps of wikis which no longer exist

A complete copy of selected Wikimedia wikis which no longer exist and so which are no longer available via the main database backup dump page. This includes, in particular, the Sept. 11 wiki.

Analytics data files

Pageview, Mediacount, Unique, and other stats.

Other files

Image tarballs, survey data and other items.

Kiwix files

Static dumps of wiki projects in OpenZim format

Dataset collection at the Data Hub (off-site)

Many additional datasets that may be of interest to researchers, users and developers can be found in this collection.

Software downloads

MediaWiki

MediaWiki is a free software wiki package written in PHP, originally for use on Wikipedia. It is now used by several other projects of the non-profit Wikimedia Foundation and by many other wikis.

Figura 4.1: Pagina principal de dumps.wikimedia.org.

Archivo	Contenido
pages-articles.xml	Este archivo contiene todos los artículos solo con su revisión mas reciente, también posee información sobre templates y descripción de archivos.
pages-logging.xml	Este archivo contiene información sobre creación o bloqueo de usuarios, cargas de imágenes, importación o movimiento de paginas, aumento de niveles de protección y eliminación de revisiones.
pages-meta-current.xml	Este archivo contiene información sobre páginas personales de usuarios y de discusión.
pages-meta-history.xml	Este archivo posee información de todos los artículos de todos los namespace de la wiki con todas las revisiones que se le realizaron a cada uno.
stub-articles.xml	Este archivo contiene la metadata todos los artículos solo con su revisión más reciente, también posee información sobre templates y descripción de archivos. No posee texto en las revisiones solo la longitud de las mismas.
stub-meta-current.xml	Este archivo contiene solo la metadata de páginas personales de usuarios y de discusión. No posee texto en las revisiones solo la longitud de las mismas.
stub-meta-history.xml	Este archivo posee metadata de todos los artículos de todos los namespace de la wiki con todas las revisiones que se le realizaron a cada uno. No posee texto en las revisiones solo la longitud de las mismas.

Tabla 4.1: WikiDumps

el archivo “pages-meta-history.xml” ya que posee el total de la información sobre los artículos y sus revisiones.

Una ventaja de esta forma de obtención de información es que obtenemos el total de los artículos de la wiki para estudiar por otro lado esta también es una desventaja por el hecho de que dependiendo que wiki se tome para estudiar el tamaño de estos archivos puede escalar rápidamente.

4.4. Funcionalidad de exportación en wikis

En las wikis que utilizan el motor MediaWiki se pueden obtener artículos en formato XML a partir de la funcionalidad Special:Export de las mismas de dos formas, una a través de la interfaz web brindada por las wikis y otra mediante solicitudes POST.

En la Figura 4.2 podemos observar la interfaz para exportar artículos de Wikipedia en inglés. En esta interfaz podemos solicitar artículos pertenecientes a una categoría o a partir del nombre del artículo, que es único y lo identifica, podemos listar los artículos que son de nuestro interés. Además nos permite seleccionar si incluir solo la revisión actual o el historial de sus revisiones, o incluir su templates o no.

La opción de realizar esta solicitud mediante una consulta POST también brinda las mismas opciones y en adición nos permite indicar a partir de qué fecha y hora queremos obtener las revisiones y la cantidad de revisiones que queremos que acompañen al artículo.

En ambos casos existe un límite sobre cuántas revisiones podemos obtener en una sola consulta, este límite es de 1000 revisiones por consulta. Mediante la interfaz web este límite siempre se aplicará a obtener solo las 1000 revisiones más antiguas, mientras que en la solicitud POST esto dependerá de la fecha y hora que le indiquemos como base.

4.5. Estructura de la información

Tanto si trabajamos con los dumps o con artículos exportados ambos se encuentran en formato xml el cual es un formato de marcado que nos permite estructurar información, por lo cual a continuación se muestra un extracto de un xml generado por la wiki. Cabe destacar que la estructura del xml es la misma sea obtenido por la funcionalidad de exportar o del “pages-meta-history.xml” .

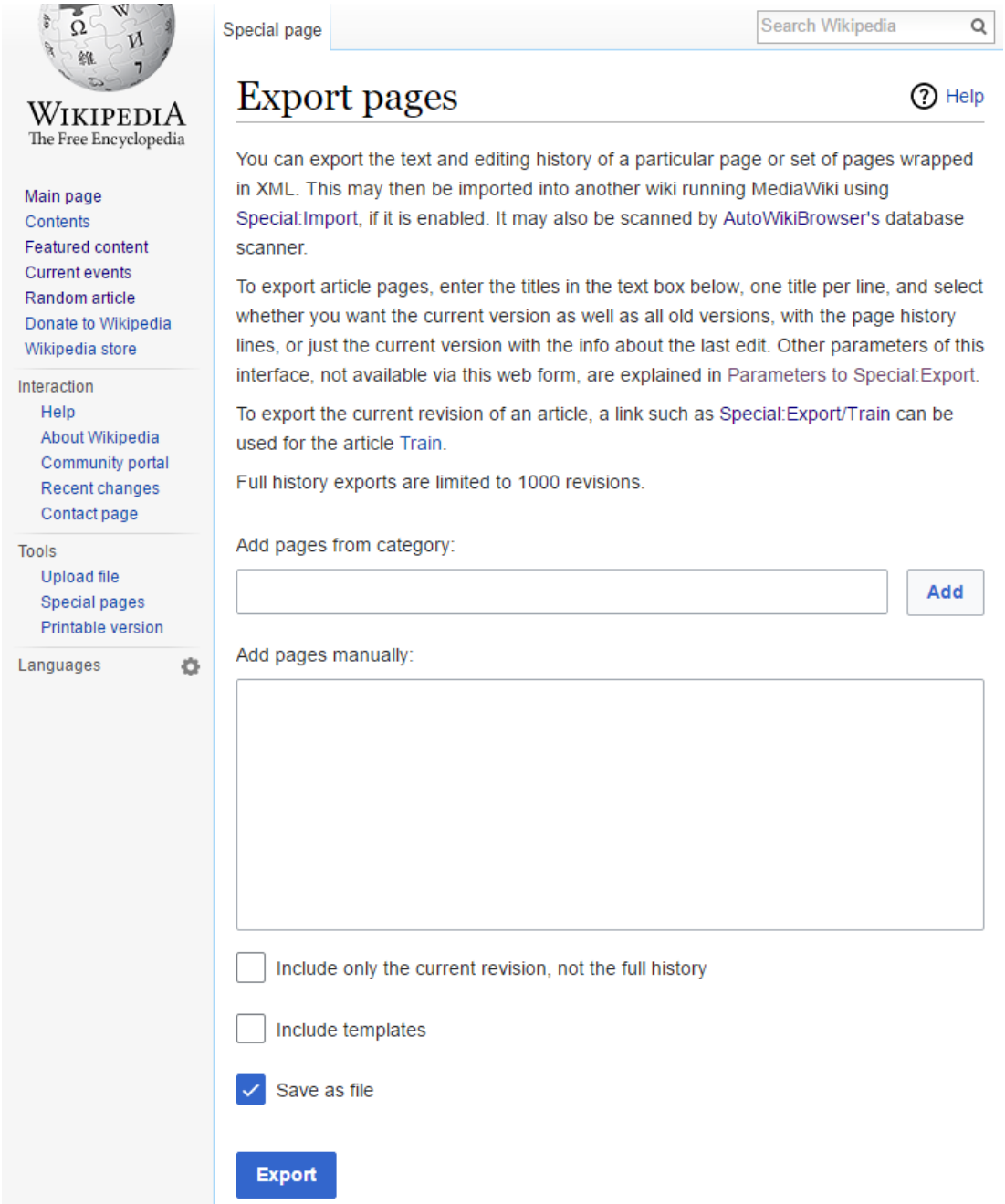


Figura 4.2: Menú para exportar páginas de Wikipedia.

```
<mediawiki xmlns="http://www.mediawiki.org/xml/export-0.10/" xmlns:xsi="http://www.w3.org
    /2001/XMLSchema-instance" xsi:schemaLocation="http://www.mediawiki.org/xml/export
    -0.10/ http://www.mediawiki.org/xml/export-0.10.xsd" version="0.10" xml:lang="en">
  <siteinfo>
    <sitename>Wikipedia</sitename>
    <dbname>enwiki</dbname>
    <base>https://en.wikipedia.org/wiki/Main_Page</base>
    <generator>MediaWiki 1.29.0-wmf.4</generator>
    <case>first-letter</case>
    <namespaces>
      <namespace key="-2" case="first-letter">Media</namespace>
      <namespace key="-1" case="first-letter">Special</namespace>
      <namespace key="0" case="first-letter" />
      <namespace key="1" case="first-letter">Talk</namespace>
      <namespace key="2" case="first-letter">User</namespace>
      <namespace key="3" case="first-letter">User talk</namespace>
      <namespace key="4" case="first-letter">Wikipedia</namespace>
      <namespace key="5" case="first-letter">Wikipedia talk</namespace>
      <namespace key="6" case="first-letter">File</namespace>
      <namespace key="7" case="first-letter">File talk</namespace>
      <namespace key="8" case="first-letter">MediaWiki</namespace>
      <namespace key="9" case="first-letter">MediaWiki talk</namespace>
      <namespace key="10" case="first-letter">Template</namespace>
      <namespace key="11" case="first-letter">Template talk</namespace>
      <namespace key="12" case="first-letter">Help</namespace>
      <namespace key="13" case="first-letter">Help talk</namespace>
      <namespace key="14" case="first-letter">Category</namespace>
      <namespace key="15" case="first-letter">Category talk</namespace>
      <namespace key="100" case="first-letter">Portal</namespace>
      <namespace key="101" case="first-letter">Portal talk</namespace>
      <namespace key="108" case="first-letter">Book</namespace>
      <namespace key="109" case="first-letter">Book talk</namespace>
      <namespace key="118" case="first-letter">Draft</namespace>
      <namespace key="119" case="first-letter">Draft talk</namespace>
      <namespace key="446" case="first-letter">Education Program</namespace>
      <namespace key="447" case="first-letter">Education Program talk</namespace>
```

```
35        <namespace key="710" case="first-letter">TimedText</namespace>
36        <namespace key="711" case="first-letter">TimedText talk</namespace>
37        <namespace key="828" case="first-letter">Module</namespace>
38        <namespace key="829" case="first-letter">Module talk</namespace>
39        <namespace key="2300" case="first-letter">Gadget</namespace>
40        <namespace key="2301" case="first-letter">Gadget talk</namespace>
41        <namespace key="2302" case="case-sensitive">Gadget definition</namespace>
42        <namespace key="2303" case="case-sensitive">Gadget definition talk</namespace>
43      </namespaces>
44    </siteinfo>
45    <page>
46      <title>Pope</title>
47      <ns>0</ns>
48      <id>23056</id>
49      <revision>
50        <id>273292</id>
51        <timestamp>2001-11-09T13:56:35Z</timestamp>
52        <contributor>
53          <username>Malcolm Farmer</username>
54          <id>135</id>
55        </contributor>
56        <minor/>
57        <comment>revert</comment>
58        <model>wikitext</model>
59        <format>text/x-wiki</format>
60        <text xml:space="preserve" bytes="12105">
61        </text>
62        <sha1>0sbfeu2zdgzgtv0d7cxj2sv4o77707j</sha1>
63      </revision>
64    </page>
65  </mediawiki>
```

Listing 4.1: Extracto de dump de Wikipedia

En el ejemplo anterior se presenta una única página con una única revisión a modo de ejemplo, también se omitió el contenido de la revisión para facilitar la lectura. La estructura del xml está dado por un nodo raíz que representa a la wiki, y está compuesto por 1+n nodos siendo estos un nodo que contiene información sobre la wiki

como el nombre y los namespaces con sus identificadores, y por n nodos representando a las páginas dentro del archivo. El nodo de las páginas a su vez está compuesto por información como el título, su namespace, su identificador y k nodos que representan cada una de sus revisiones con la información propia de las mismas.

4.6. Spring y Hibernate

Para el desarrollo del prototipo se utilizó el Framework Spring integrado con el ORM Hibernate. Se decidió la utilización de estas herramientas por el conocimiento previo del autor en las mismas y por las características que se mencionan a continuación.

El framework Spring por su parte nos permite desarrollar el prototipo web de forma rápida y dinámica simplificando el punto de partida del desarrollo. También otorga herramientas como soporte MVC, inyección de dependencias e integración con múltiples formas de persistencia en base de datos.

Por otro lado el ORM Hibernate permite trabajar de forma orientada a objetos en lo relacionado a persistencia. Además de que nos permite abstraernos del modelo de la base de datos a la hora de trabajar o lo que fue de mayor interés para nosotros abstraernos sobre que motor de base de datos utilizar. Por último es importante resaltar que es fácilmente integrable con el framework Spring.

Durante este trabajo se trabajaron con dos tipos de bases de datos para realizar las pruebas, una base de datos tradicional utilizando el motor mysql y una base de datos en memoria para mejorar la velocidad del prototipo utilizando el motor H2.

4.7. Carga de wiki completa

Podemos obtener el total de la información de los artículos de una wiki a partir de sus dumps como se explicó anteriormente. Para ello el prototipo nos permite a partir de un archivo configurar el directorio en el que se ubica el dump que deseamos cargar. Dicho archivo puede ubicarse en:

- Desplegado en tomcat: /Tomcat/webapps/WikiWebTest/WEB-INF/clases/historyPath.properti
- En el proyecto eclipse: /src/main/resources/historyPath.properties

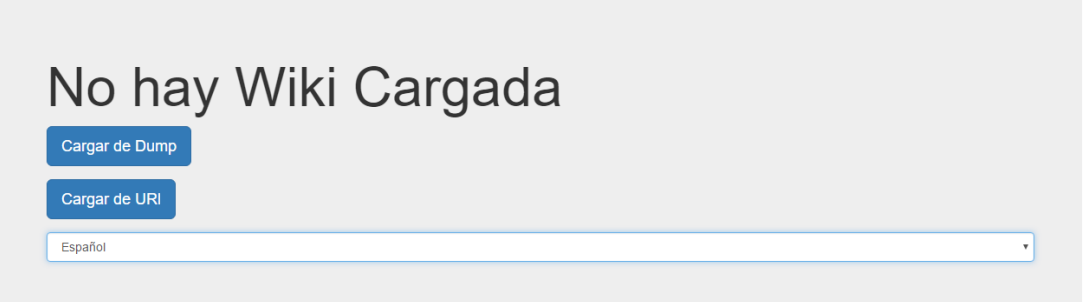


Figura 4.3: Menú de WikiWebTest.

Una vez localizado el archivo debemos modificarlo indicando el directorio donde se encuentra el archivo y utilizando caracteres de escape de requerirse. Para mas información de las configuraciones de este archivo referirse al Apéndice A

Luego ingresando a la aplicación, $http://localhost:8080/WikiWebTest/$, como se puede observar en la Figura 4.3 en caso de que no haya ninguna otra wiki cargada en la base de datos disponemos de la opción "Cargar de Dump" la cual realiza nos dará acceso al menú que se observa en la Figura 4.4. Desde este menú se nos permite seleccionar la clave con la que se configuro el dump que deseamos cargar y también si lo deseamos solicitar la generación automática de las estadísticas o la carga dinámica descargando las revisiones directamente de Wikipedia. También nos permite limitar los resultados finales a un numero que seleccionemos de artículos los cuales se seleccionaran del total utilizando la clase Random"del paquete java.util, ademas para permitir resultados repetibles permite la configuración de una semilla que permita obtener de forma regular los mismos resultados. Es importante mencionar que la función de selección aleatoria y de configuración de semilla solo se encuentra habilitadas para cargas dinámicas

Además aunque las estadísticas generales de la wiki no se ven restringidas por la forma de carga es cierto que solo con esta forma de carga en particular cobran real relevancia dado que no poseen sesgo de ningún tipo.

4.8. Carga de artículos específicos

Dado el tamaño de la información contenido en un dump de una wiki y el tiempo que puede llevarnos al procesarlo completo se decidió implementar una funcionalidad que permita seleccionar artículos de interés para su estudio.

Figura 4.4: Menú carga de dump de WikiWebTest.

Figura 4.5: Menú carga desde URI de WikiWebTest.

Para esto se debe seleccionar la opción "Cargar de URI" que se observa en la Figura 4.3, esta opción nos brinda acceso a la interfaz que se muestra en la Figura 4.5. En esta interfaz podemos listar los artículos de nuestro interés para luego iniciar la descarga. También al igual que en la carga de dump se permite seleccionar la opción de realizar la carga automática de las estadísticas

En este método se trabaja a partir de la uri del artículo de interés a fin de obtener el título exacto del mismo con el que se lo identifica. Por ejemplo para obtener el artículo de Barack Obama utilizaremos la parte resaltada de la url que se muestra a continuación.

$https://en.wikipedia.org/wiki/Barack\_Obama$

Para la carga de artículos específicos el prototipo se encuentra limitado a artículos de la Wikipedia en inglés, esta limitación es debido a la forma de acceso a la API la cual es mediante solicitudes POST ademas de que en este trabajo se estudia particularmente Wikipedia en su versión en inglés.

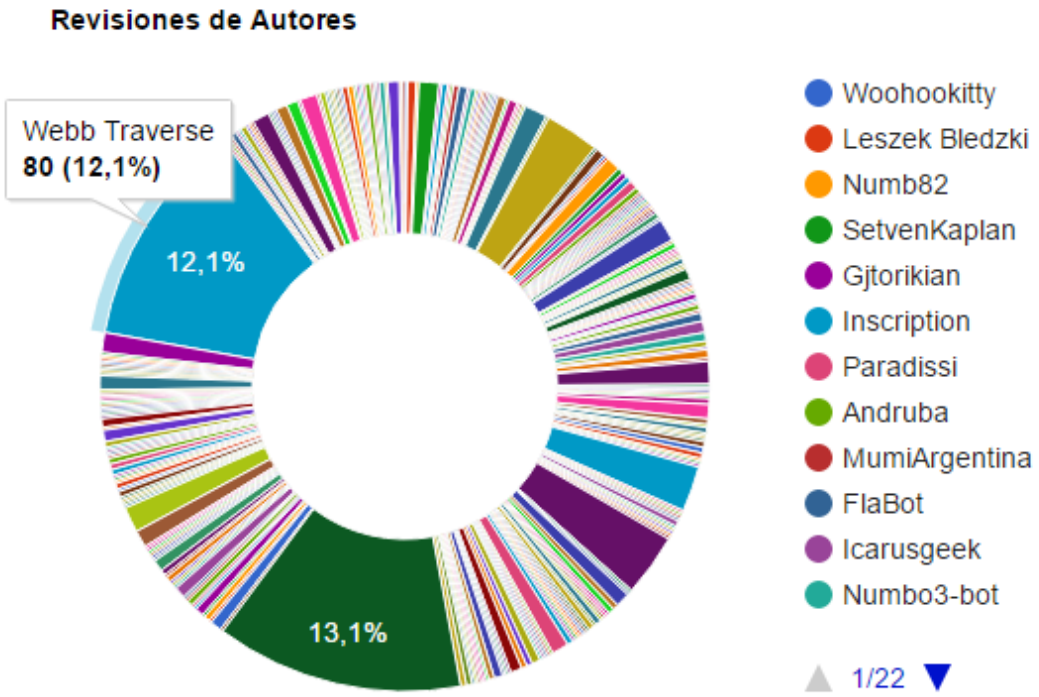


Figura 4.6: Revisiones de autores de un artículo.

4.9. Estadísticas de artículos

A continuación se presenta la información que obtenemos de la página de estadísticas de un artículo. En la figura 4.6 podemos observar un gráfico que representa el porcentaje de revisiones realizada por cada usuario junto con un listado paginado de los usuarios que realizaron ediciones. También puede observarse que al posicionarse sobre alguna porción del gráfico se nos brinda información más detallada como por ejemplo la cantidad exacta de revisiones realizadas por dicho autor. Este gráfico fue realizado utilizando el resultado de aplicar la métrica *PercentOfRevisionByAuthor*, que a su ves utiliza las métricas *PercentOfAuthorRevisions* y *#RevisionsOfAuthor*, a la información de las revisiones de un articulo.

Continuando con la información obtenida de las estadísticas de revisiones obtenemos el gráfico que se observa en al Figura 4.7 que representa la cantidad de revisiones que se hicieron cada día, el mismo dispone de dos partes, la superior es un control que permite delimitar una porción de la información del total la cual se representará en la

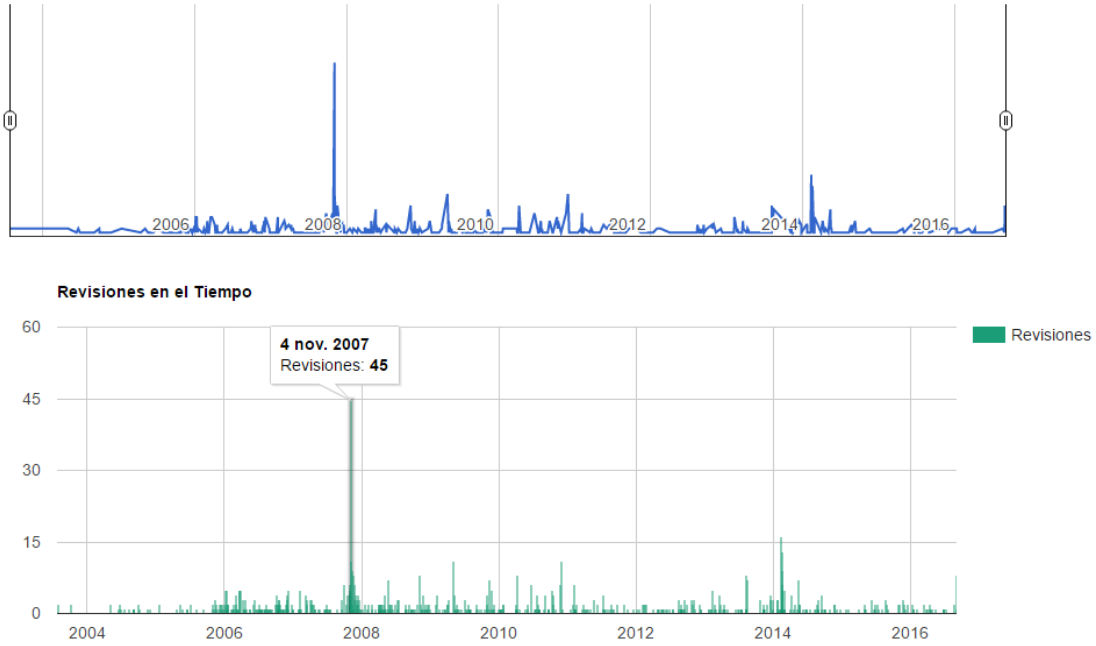


Figura 4.7: Revisiones en el tiempo de un artículo.

parte inferior, en el segundo gráfico de la Figura 4.8 podemos ver esta funcionalidad en acción acotando el periodo de tiempo visualizado. También seleccionando parte del gráfico podemos obtener información específica de la cantidad de revisiones que se realizaron en una determinada fecha. La forma correcta de interpretar el gráfico es tomando el eje X como el periodo de tiempo desde que se creo el articulo a la actualidad, y el eje Y como el indicador del numero revisiones realizadas por lo cual cuanto mas elevadas son las columnas del gráfico mas revisiones se realizaron en un día en particular. Este gráfico fue realizado utilizando el resultado de aplicar la métrica *RevisionPerDay*, que a su ves utiliza la métrica *RevisionsOfDay*, a la información de las revisiones de un articulo.

Finalizando con las estadísticas de revisiones se presenta un gráfico en la Figura 4.9 en el cual se muestra el tamaño en bytes de las revisiones en el tiempo, se eligió esta representación por la relación de equivalencia entre un carácter y un byte. Este gráfico posee una visión de zoom al igual que en los gráficos anteriormente descritos, y también permite al seleccionar parte del gráfico obtener información adicional como el tamaño exacto en bytes. Es importante resaltar que en los gráficos pueden observarse líneas en blanco, las mismas no representan necesariamente que el contenido se haya reducido drásticamente a 0 sino que es una limitación del graficador el cual se alimenta de la revisiones para generar el gráfico y por lo cual de no existir revisiones en esa fecha

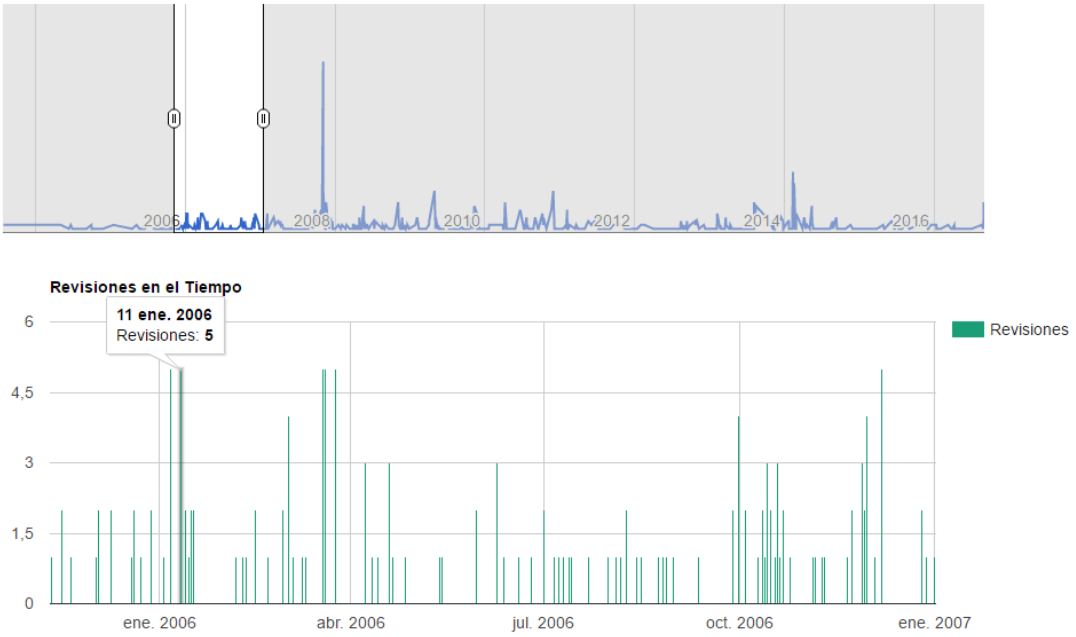


Figura 4.8: Revisiones en el tiempo de un artículo con zoom.

simplemente representa la fecha como si el contenido fuera de 0 bytes. En este gráfico el eje X representa el periodo de tiempo desde que se creo el articulo a la actualidad, y el eje Y como el indicador del numero bytes contenidos en el articulo por lo cual cuanto mas elevadas son las columnas del gráfico mayor es el tamaño del articulo para esa fecha. Este gráfico fue realizado utilizando el resultado de aplicar la métrica *BytesOfRevisionsPerDay* a la información de las revisiones de un articulo.

4.10. Estadísticas de estilos de artículos

Continuando con las estadísticas se presentan las relacionadas con la información de los estilos en el primer gráfico que aparece nos brinda un controlador para seleccionar cuáles tipos de estilos queremos visualizar en el gráfico de tal forma de poder visualizarlos de forma individual en la Figura 4.10 o grupal en la Figura 4.11, en este último caso a la derecha se presenta un listado indicando el color con el que se representa al estilo en el gráfico llegando a convertirse en un listado paginado según la cantidad de estilos presentados. En ambos gráficos posicionándonos sobre el mismo podemos obtener información adicional de la cantidad de apariciones de un estilo en particular. La lectura de este gráfico es con el eje X representando el periodo de tiempo desde que se

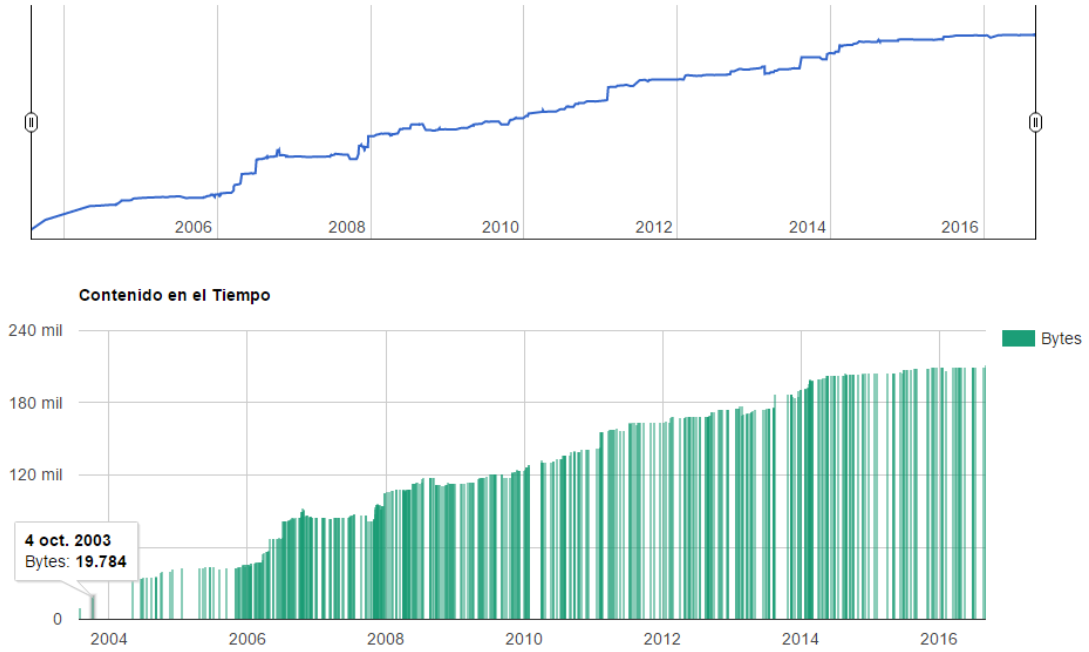


Figura 4.9: Contenido en el tiempo de la un artículo.

creo el articulo a la actualidad, y el eje Y como el indicador del numero de aplicaciones del estilo en el articulo por lo cual cuanto mas elevados sean alguno de los trazos que representan a cada tipo de estilo seleccionado mayor sera la cantidad de sus aplicaciones para una fecha determinada. Este gráfico fue realizado utilizando el resultado de aplicar la métrica *#OcurrencesOfStyle* a la información de markup extraída de el texto de cada revisión del articulo.

Por último también podemos observar un gráfico en la Figura 4.12 en el cual se visualiza la cantidad total de estilos por revisión, al igual que en los gráficos anteriores al posicionarse sobre el mismo se obtiene información sobre la cantidad exacta de estilos. Además el gráfico provee del mismo tipo de funcionalidad de zoom utilizado en gráficos anteriores. Este gráfico posee una forma de lectura que continua con la de los gráficos anteriores siendo el eje X el periodo de tiempo mientras el eje Y representa el valor a medir en esta ocasión la sumatoria de todos los estilos de cada revisión, por lo cual en que las columnas aumentan de tamaño también aumenta el numero de estilos aplicados en el articulo. Este gráfico fue realizado utilizando el resultado de aplicar la métrica *StylesOfRevisionsPerDay*, que a su ves utiliza la métrica *StylesInText*, a la información de markup extraída de el texto de cada revisión del articulo.

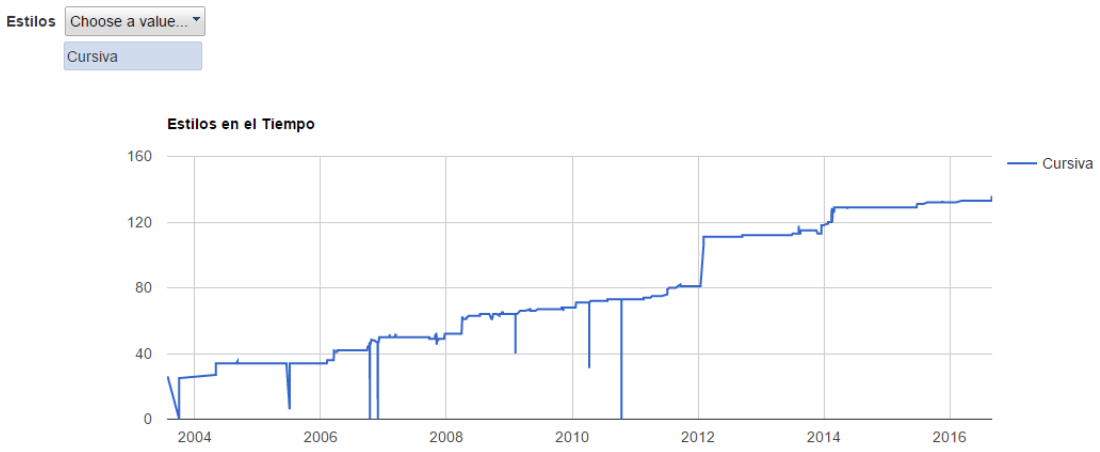


Figura 4.10: Estilos en el tiempo de un artículo. Solo Cursiva.

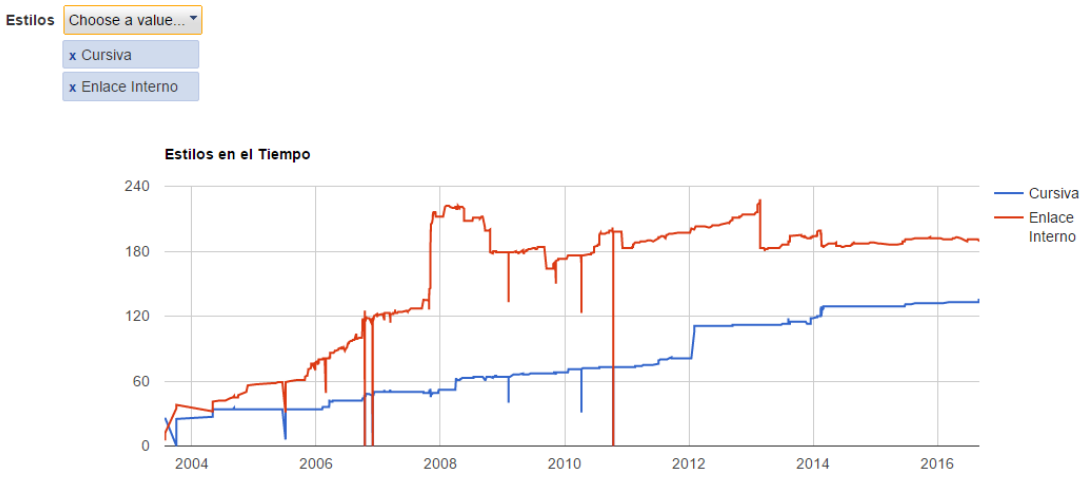


Figura 4.11: Estilos en el tiempo de un artículo. Estilos Cursiva y Enlace Interno.

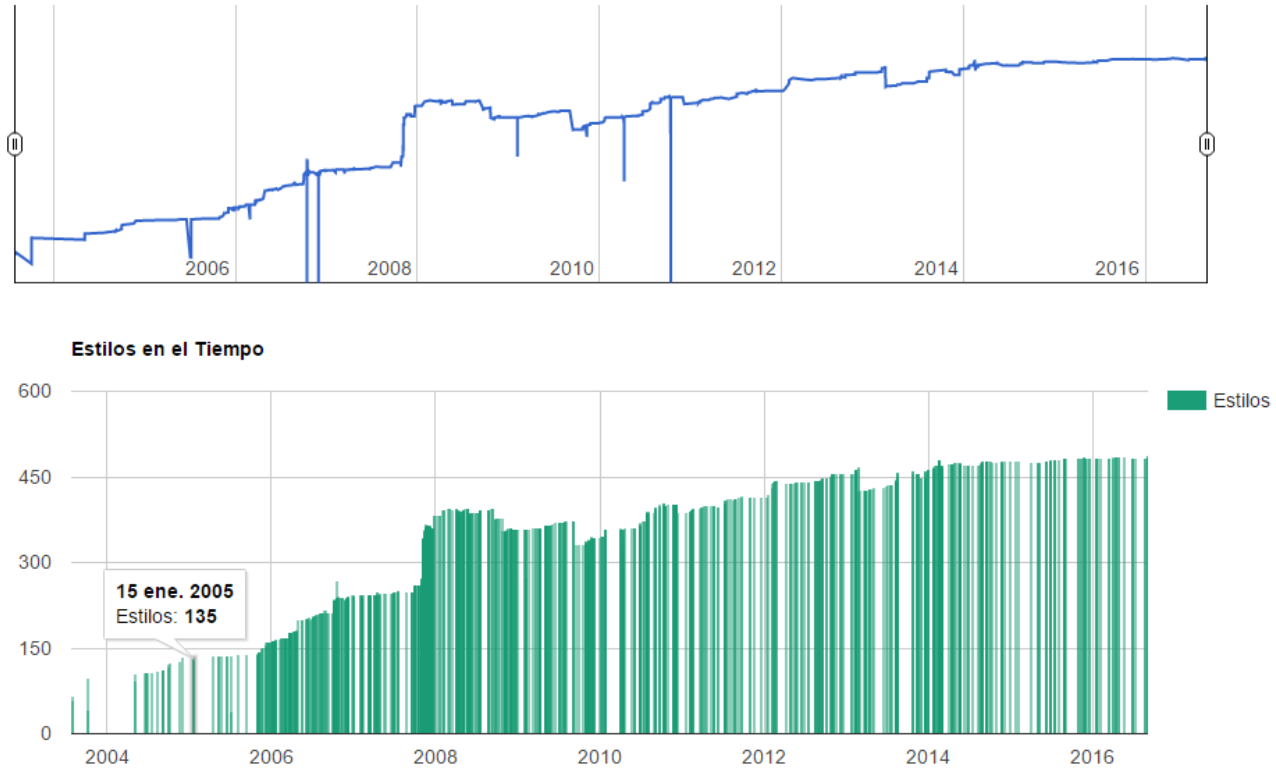


Figura 4.12: Estilos en el tiempo de un artículo. Estilos acumulados.

4.11. Generación de json

Por último este prototipo brinda la funcionalidad de exportar toda la información generada de los artículos generando un archivo json por cada artículo. Para ello una vez cargada la wiki desde el menú principal que podemos observar en la Figura 4.13 el botón "Exportar Estadísticas" que nos permitirá llevar a cabo dicha acción depositando los resultados en la ubicación previamente configurada en el archivo historyPath.properties como se indica en el Apéndice A. Estos json serán utilizados en la segunda parte de este trabajo donde además se explicará el contenido de los mismos con detalle.

Figura 4.13: Menú de WikiWebTest. Con una wiki cargada.

Capítulo 5

Evaluación y resultados

5.1. Introducción

En esta sección se define el conjunto de datos sobre el que se realizará la evaluación y se explica el método de evaluación empleado con detalles de cómo inicializa prototipo. Posteriormente se presentarán los resultados tal como se obtuvieron del prototipo y luego se presenta un análisis de los casos más relevantes del mismo.

5.2. Conjunto de datos

Como se mencionó anteriormente en este trabajo se utilizaran wikis de la familia de MediaWiki, en particular para realizar la evaluación se utilizaron las siguientes páginas extraídas de la versión en inglés de Wikipedia:

Título	Fecha de primera revisión obtenida	Fecha de última revisión obtenida
Julio Cortázar	29 de julio del 2003	29 de agosto de 2016
Pope	9 de noviembre del 2001	18 de septiembre de 2016
Johnny Depp	14 de agosto del 2002	23 de septiembre de 2016
Barack Obama	18 de marzo del 2004	22 de septiembre de 2016

Tabla 5.1: Conjunto de datos

5.3. Método de evaluación

Para realizar esta evaluación primero se creó una base de datos h2 en la cual el prototipo pudiese almacenar la información obtenida y se procedió a la puesta en producción del prototipo en un servidor Tomcat donde luego se configuró información de la localización y datos de acceso para la base de datos.

Con el prototipo ya configurado se accedió la interfaz web del mismo desde la cual se le indico al prototipo que baje las páginas listadas anteriormente y además que pre calcule todas las estadísticas. A partir de ese momento el prototipo comienza a solicitar a Wikipedia la página solicitada con todas sus revisiones y las categorías a la que esta haya pertenecido, este proceso lo lleva a cabo comunicándose a través de la API "Special:Export" como ya se explicó anteriormente.

Finalizada la descarga el prototipo comienza el cómputo de estadísticas y diferencias entre revisiones de las páginas cargadas a partir de las métricas ya definidas. A partir de que finaliza accediendo a la pantalla principal del prototipo obtenemos acceso al listado de páginas y sus respectivos listados de revisiones desde los cuales podemos ver las diferencias de las mismas y también obtenemos acceso a las estadísticas calculadas anteriormente que se representan a forma de gráficos en la interfaz web.

5.4. Resultados

A continuación se presentan los resultados de la ejecución de la evaluación descrita anteriormente, en primera instancia se mostrarán los resultados de las estadísticas de revisiones y en segunda instancia las estadísticas de los estilos de la página.

5.4.1. Estadísticas de revisiones

En primera instancia por cada página se obtuvo el número de revisiones contenidas en el conjunto de datos evaluados, obteniendo :

La siguiente información que obtuvimos fueron los gráficos representan el porcentaje de revisiones realizada por cada usuario en cada articulo, a continuación se presentan los 4 gráficos resultantes de aplicar la métrica *PercentOfRevisionByAuthor*, que a su ves utiliza las métricas *PercentOfAuthorRevisions* y *#RevisionsOfAuthor*, a la información de las revisiones de los artículos Con estos gráficos podemos por lo tanto saber que autores y en que medida contribuyeron al articulo.

Título	Periodo de evaluación	Cantidad de revisiones
Julio Cortázar	29/07/2003 - 29/08/2016	1009
Pope	09/11/2001 - 18/09/2016	5374
Johnny Depp	14/08/2002 - 23/09/2016	9638
Barack Obama	18/03/2004 - 22/09/2016	25205

Tabla 5.2: Datos obtenidos

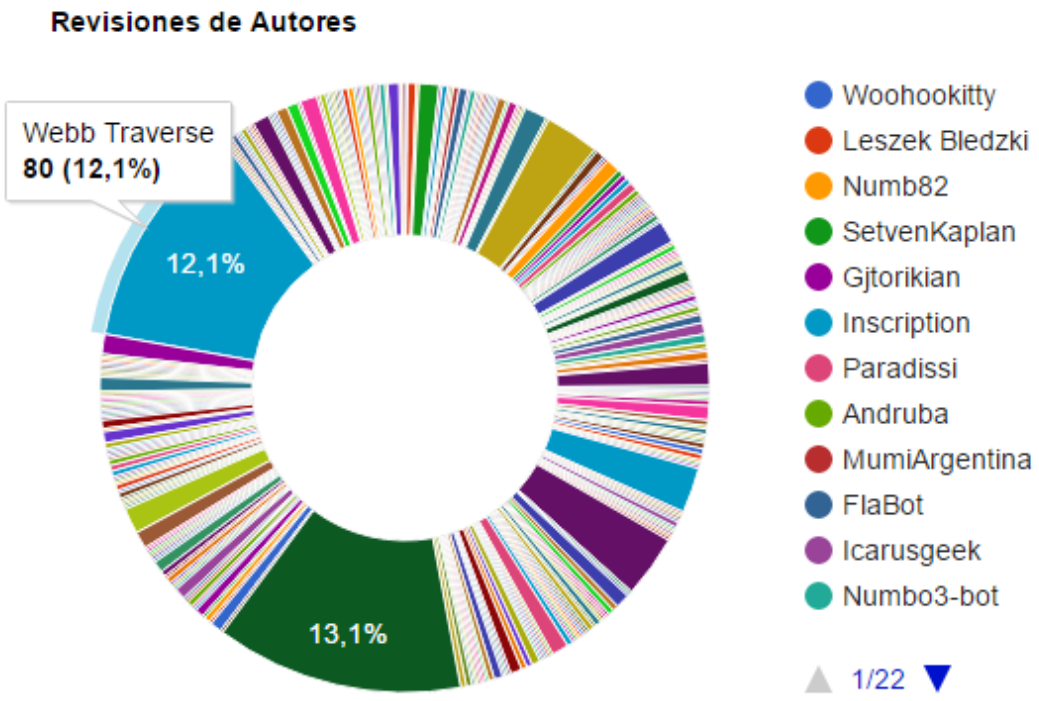


Figura 5.1: Revisiones de autores de la página Julio Cortázar.

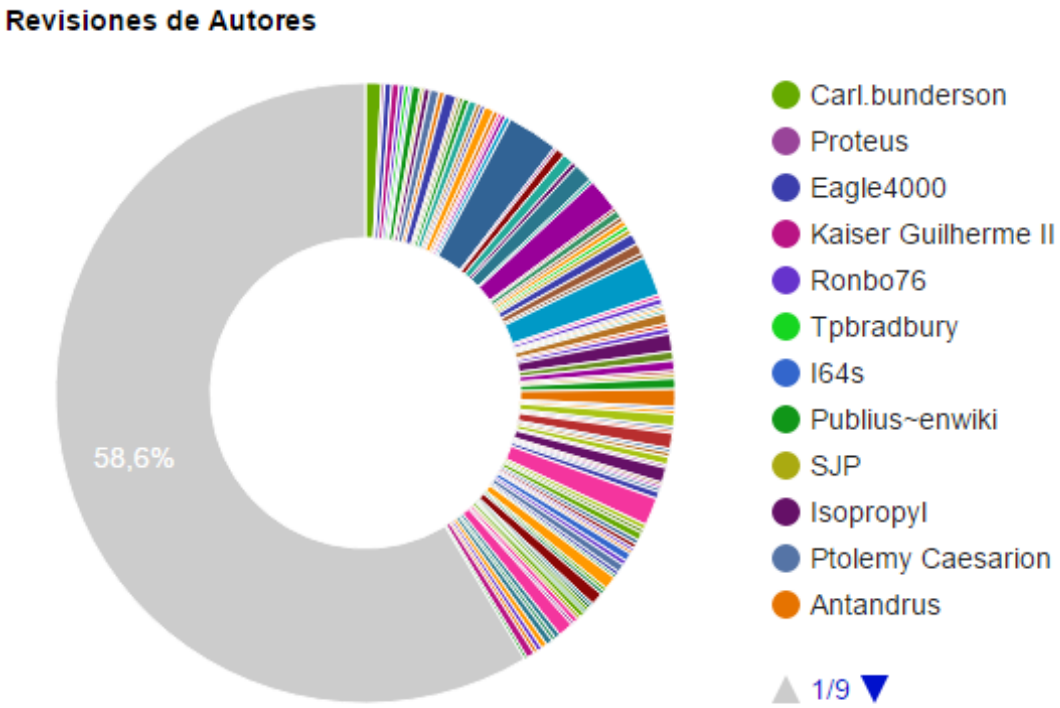


Figura 5.2: Revisiones de autores de la página Pope.

En la Figura 5.1 podemos observar el porcentaje de de revisiones realizadas por cada autor. Particularmente en el articulo de Julio Cortázar los dos usuarios que realizaron la mayor cantidad de revisiones son el usuario "Webb Traverseçon el 12,1 % de las revisiones como se observa en el gráfico y el usuario "850 C"que realizo el 13,1 % de revisiones al articulo.

En la Figura 5.2 podemos observar el porcentaje de de revisiones realizadas por cada autor. Particularmente en el articulo del Papa observamos un porcentaje del 58,6 % en gris, este porcentaje representa la cantidad de revisiones realizadas por usuarios anónimos siendo por lo tanto mayoritarias dado que el autor que tiene mayor porcentaje de revisiones realizadas en comparación solo realizo el 2,7 % de revisiones del articulo.

En la Figura 5.3 podemos observar el porcentaje de de revisiones realizadas por cada autor. Particularmente en el articulo de Johnny Depp observamos un porcentaje del 57,1 % en gris, este porcentaje representa la cantidad de revisiones realizadas por usuarios anónimos siendo por lo tanto mayoritarias como sucede en el articulo del Papa, aun que en esta ocasión el autor que tiene mayor porcentaje de revisiones realizadas

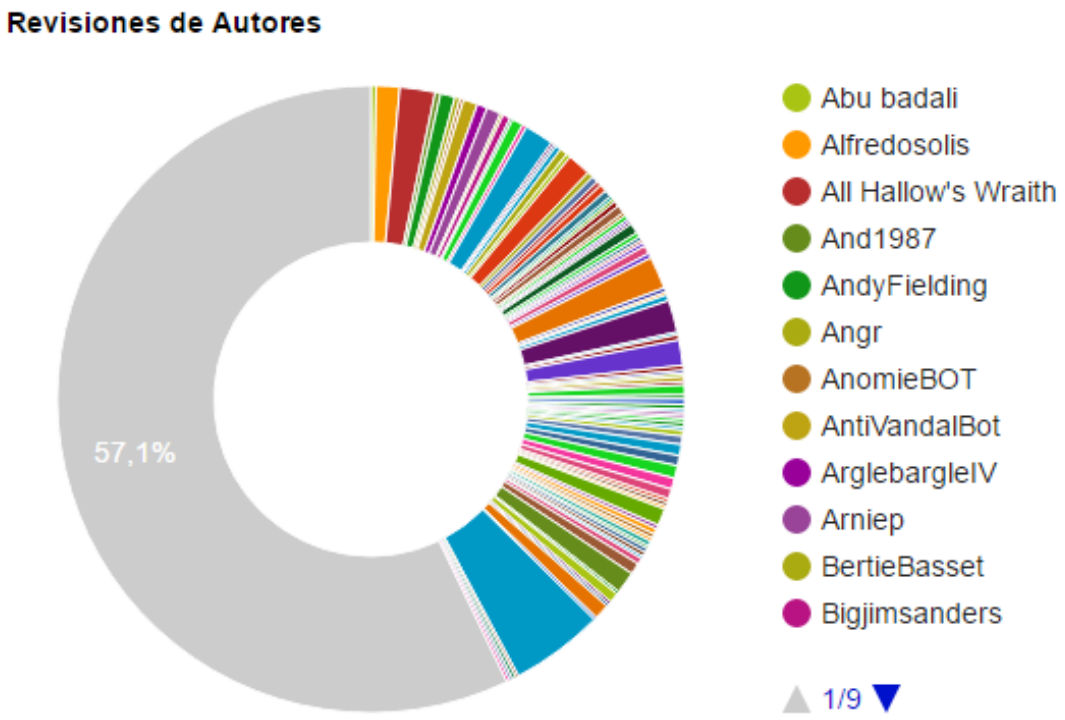


Figura 5.3: Revisiones de autores de la pagina Johnny Depp.

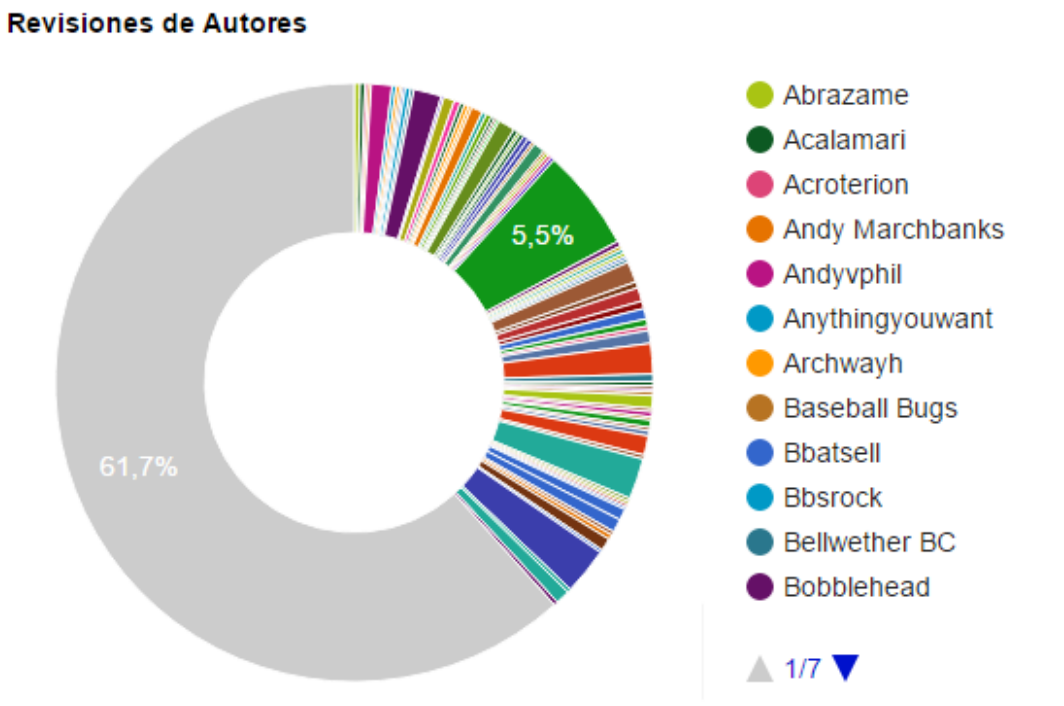


Figura 5.4: Revisiones de autores de la página Barack Obama.

realizo el 4,8 % de revisiones del articulo.

En la Figura 5.4 podemos observar el porcentaje de de revisiones realizadas por cada autor. Particularmente en el articulo de Barack Obama observamos un porcentaje del 61,7 % en gris, este porcentaje representa la cantidad de revisiones realizadas por usuarios anónimos siendo por lo tanto mayoritarias, aun que las ediciones realizada por usuarios anónimos es la mayor de todos los artículos mencionados anteriormente también encontramos que en esta ocasión el autor que tiene mayor porcentaje de revisiones realizadas realizo el 5,5 % de revisiones del articulo siendo también el mayor porcentaje realizado por un único autor en los artículos que se presentan grandes porcentajes de revisiones anónimas

Continuando con la información obtenida de las estadísticas de revisiones obtenemos los gráficos que representan la cantidad de revisiones que se hicieron cada día para cada artículo. Estos gráficos resultantes de aplicar la métrica *RevisionPerDay*, que a su ves utiliza la métrica *RevisionsOfDay*, a la información de las revisiones de los artículos nos permiten poder conocer el nivel de actividad en la edición de los artículos a lo largo del tiempo.

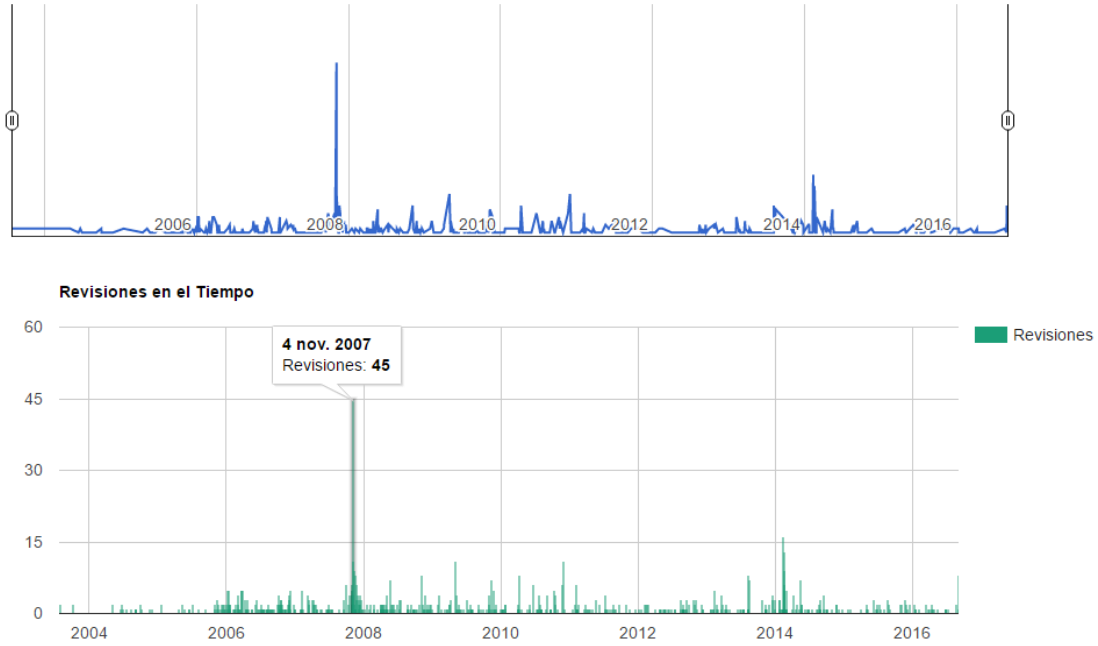


Figura 5.5: Revisiones en el tiempo de la página Julio Cortázar.

En la Figura 5.5 se muestra la actividad día a día en el articulo de Julio Cortázar. Como podemos observar el gráfico muestra un incremento de actividad aproximadamente desde el 2006 y pico de actividad el 4 de noviembre del 2007 día en el cual se realizaron 45 revisiones, la mayor cantidad de revisiones realizadas en un día en el articulo.

En la Figura 5.6 se muestra la actividad día a día en el articulo del Papa. Como podemos observar el gráfico muestra un periodo de mayor actividad aproximadamente desde el 2005 hasta principios del 2010 y posee dos puntos de mayor actividad en abril del 2005 y marzo del 2013.

En la Figura 5.7 se muestra la actividad día a día en el articulo de Johnny Depp. Como podemos observar el gráfico muestra un periodo de mayor actividad aproximadamente desde el 2005 hasta 2010 aun que a partir del incremento de actividad del 2005 se presentan durante el resto de la historia del articulo múltiples picos del nivel de actividad del articulo.

En la Figura 5.8 se muestra la actividad día a día en el articulo de Barack Obama. Como podemos observar el gráfico muestra un periodo de mayor actividad aproximada-

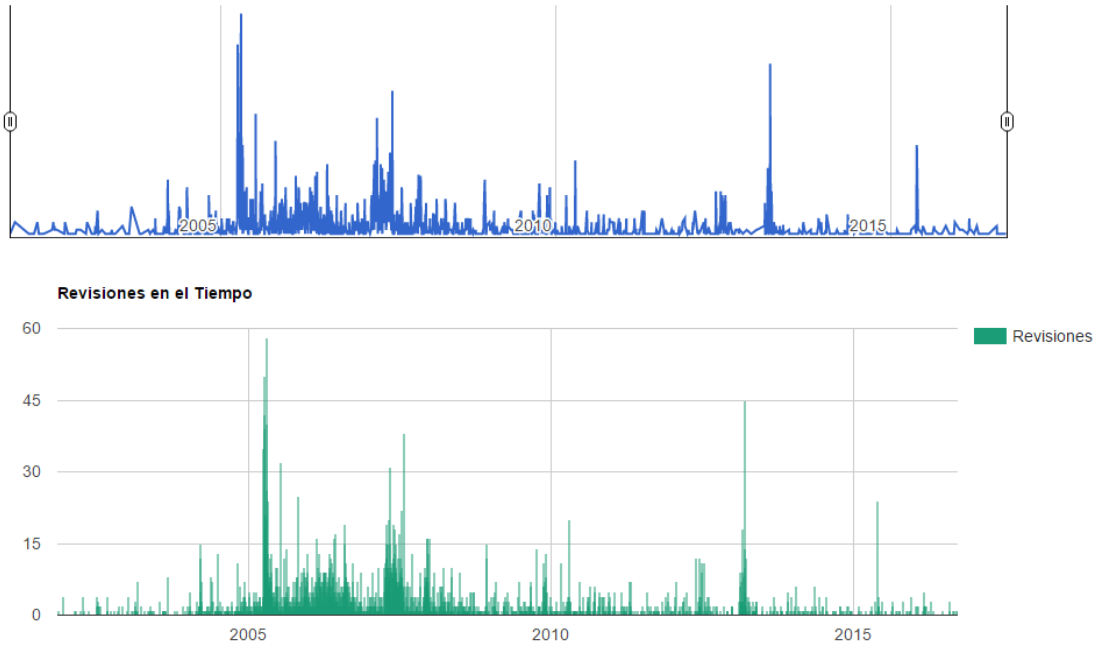


Figura 5.6: Revisiones en el tiempo de la página Pope.

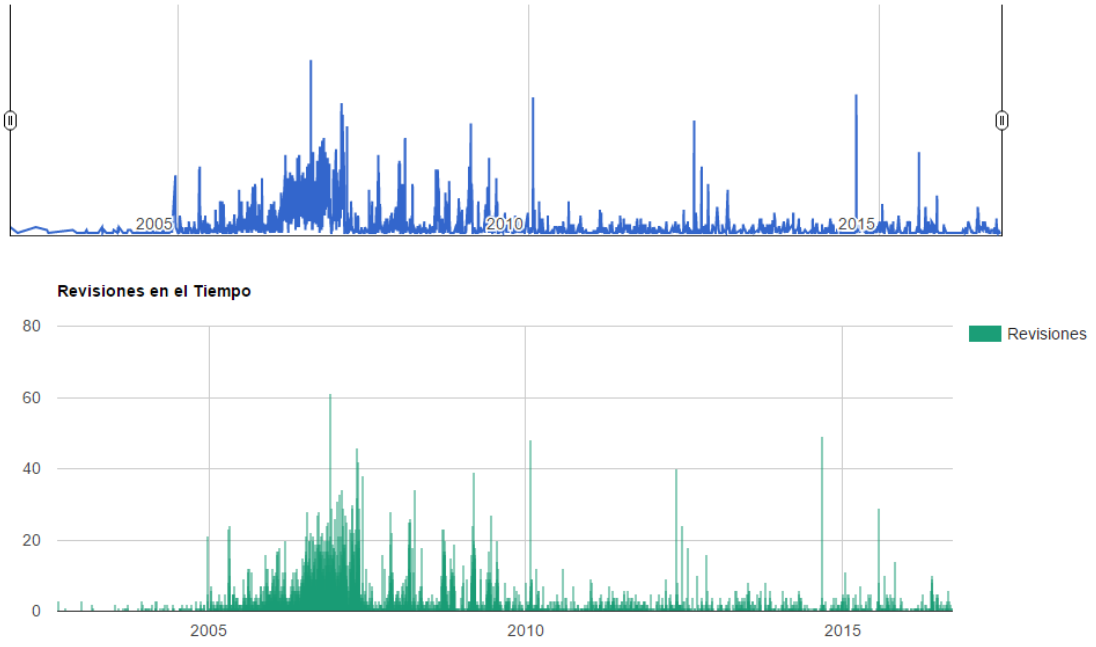


Figura 5.7: Revisiones en el tiempo de la pagina Johnny Depp.

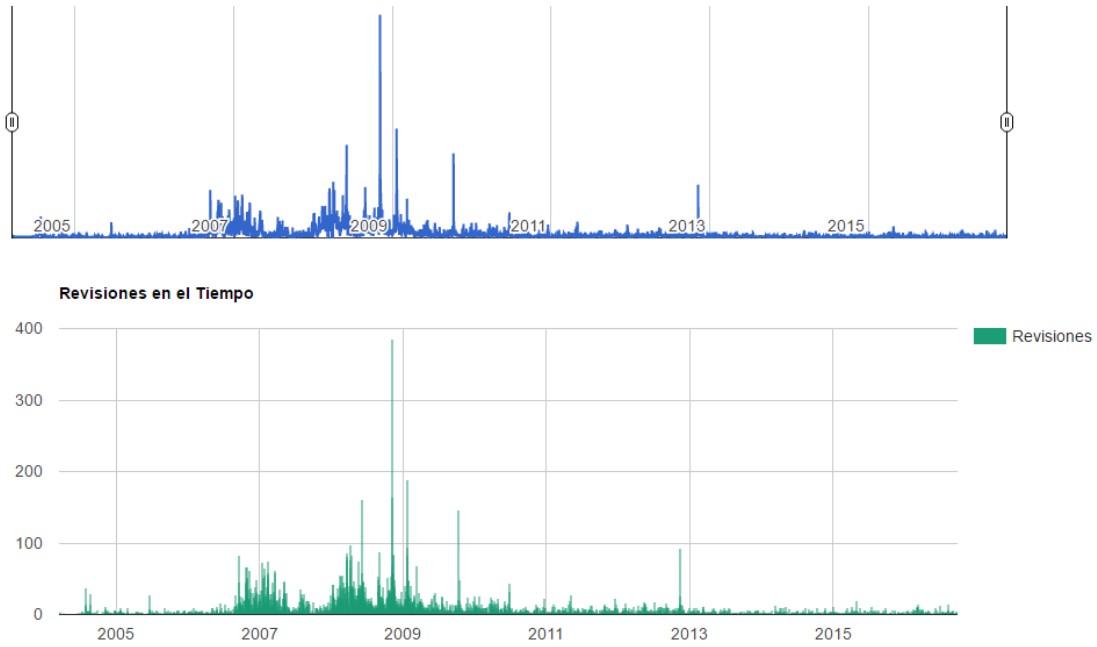


Figura 5.8: Revisiones en el tiempo de la página Barack Obama.

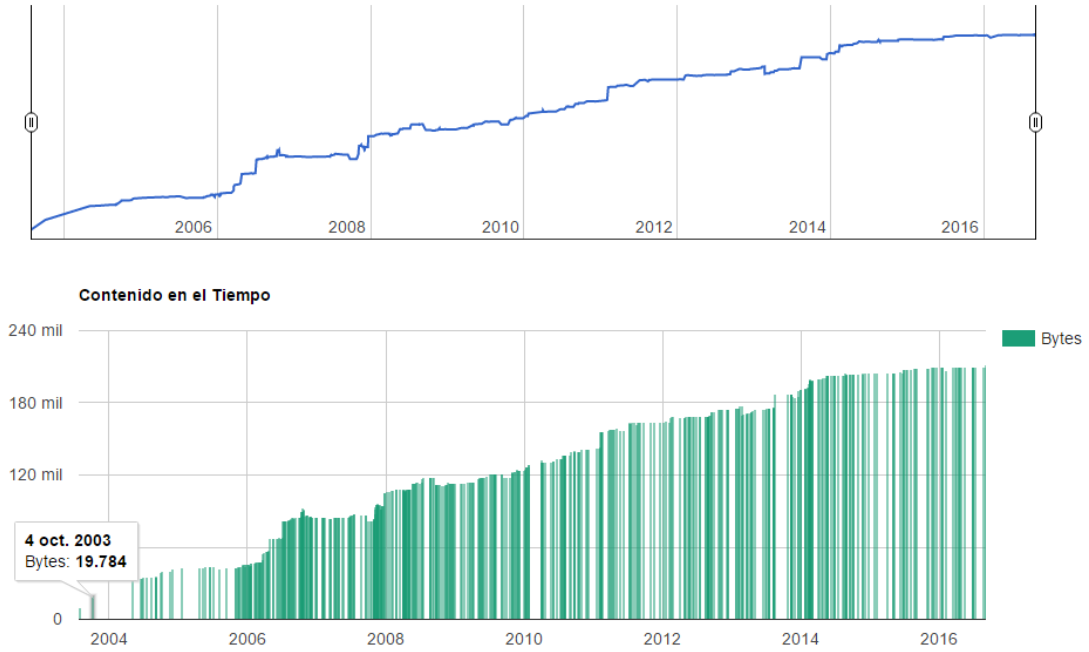


Figura 5.9: Contenido en el tiempo de la página Julio Cortázar.

mente desde finales del 2006 hasta 2010 y su fecha de mayor actividad es en noviembre del 2008.

Finalizando con las estadísticas de revisiones se presentan los gráficos en los cuales se muestran el tamaño en bytes de las revisiones en el tiempo para cada artículo. Estos gráficos fueron generados utilizando la métrica *BytesOfRevisionsPerDay* aplicándola a la información de las revisiones de los artículos. Estos gráficos nos permiten saber cuales fueron el crecimiento y decrecimiento respecto al tamaño de las revisiones de los artículos

En la Figura 5.9 se muestra la evolución del articulo de Julio Cortázar respecto al tamaño en bytes. En este gráfico podemos observar que el crecimiento del articulo comienza con mayor intensidad a mediados del 2006, luego el tamaño del articulo se estabiliza desde 2014 a la actualidad.

En la Figura 5.10 se muestra la evolución del articulo del Papa respecto al tamaño en bytes. En este gráfico podemos observar que el crecimiento del articulo se ve interrumpido a mediados del 2002 por un decrecimiento que se mantiene hasta finales del 2003 donde comienza con mayor intensidad una etapa de crecimiento del articulo, aun que con casos de decrecimientos abruptos aislados, luego el tamaño del articulo se estabiliza desde 2012 a la actualidad.

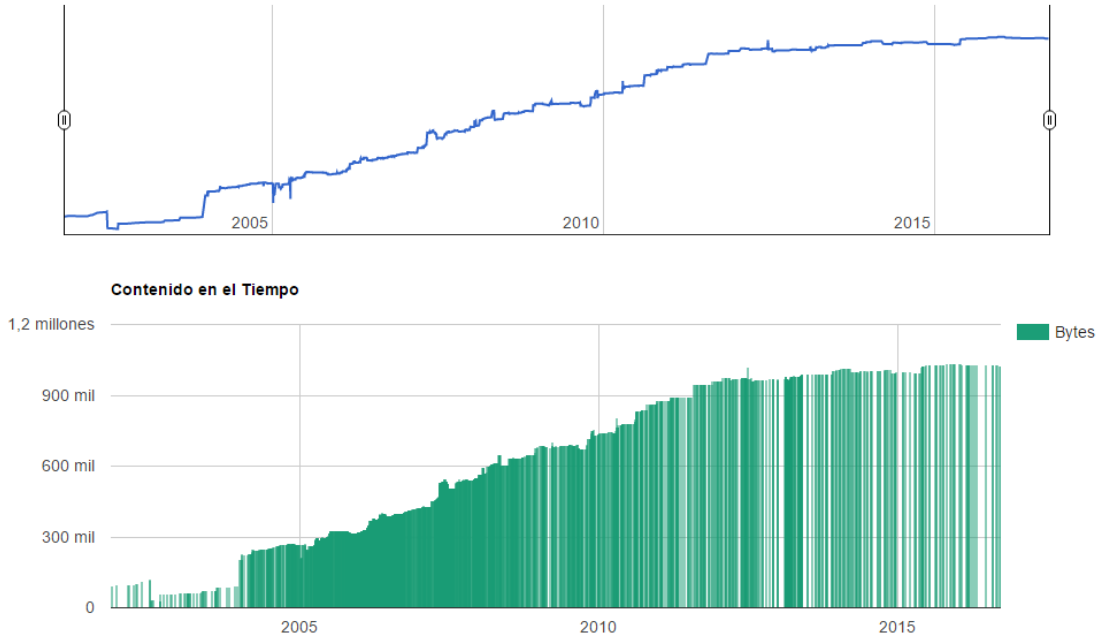


Figura 5.10: Contenido en el tiempo de la página Pope.

En la Figura 5.11 se muestra la evolución del articulo de Johnny Depp respecto al tamaño en bytes. En este gráfico podemos observar que el crecimiento del articulo es regular desde un comienzo pero a finales del 2005 presenta un incremento aislado del tamaño del articulo que duplica hasta al máximo tamaño alcanzado por el mismo en el resto de su historia.

En la Figura 5.12 se muestra la evolución del articulo de Barack Obama respecto al tamaño en bytes. En este gráfico podemos observar que el crecimiento del articulo obtiene el mayor impulso a finales del 2006 y principios del 2007 y continua en crecimiento hasta la actualidad.

5.4.2. Estadísticas de estilos

De estas estadísticas el primer gráfico que obtenemos nos permite la selección de los estilos de interés, el gráfico esta generado a partir de la métrica $\#OcurrencesOfStyle$ que es aplicada a la información de markup extraída de el texto de cada revisión de los artículos. Este gráfico nos permite analizar de forma especifica los cambios de estilos aplicados entre dos o mas revisiones de un articulo en particular. Dada la gran cantidad de combinaciones posibles de los mismos este gráfico sólo será mostrado para casos

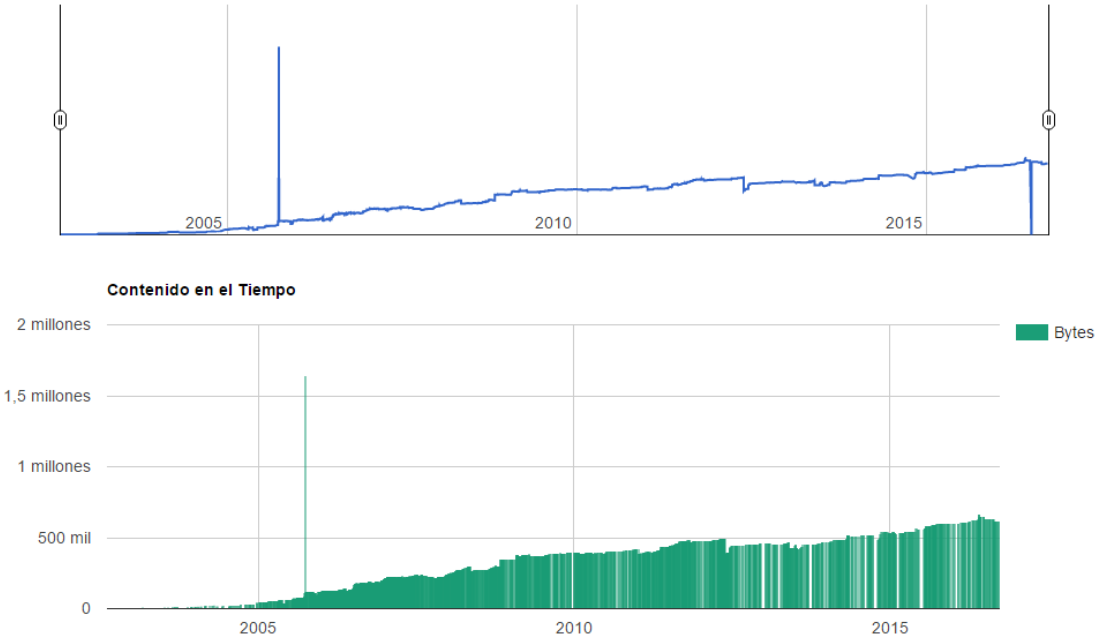


Figura 5.11: Contenido en el tiempo de la pagina Johnny Depp.

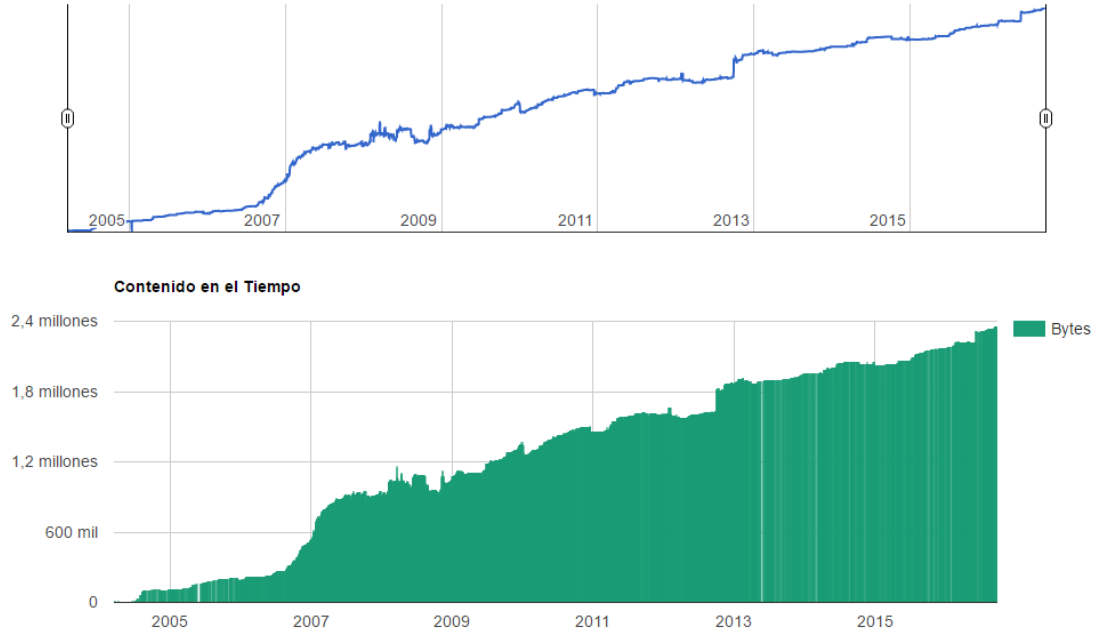


Figura 5.12: Contenido en el tiempo de la página Barack Obama.

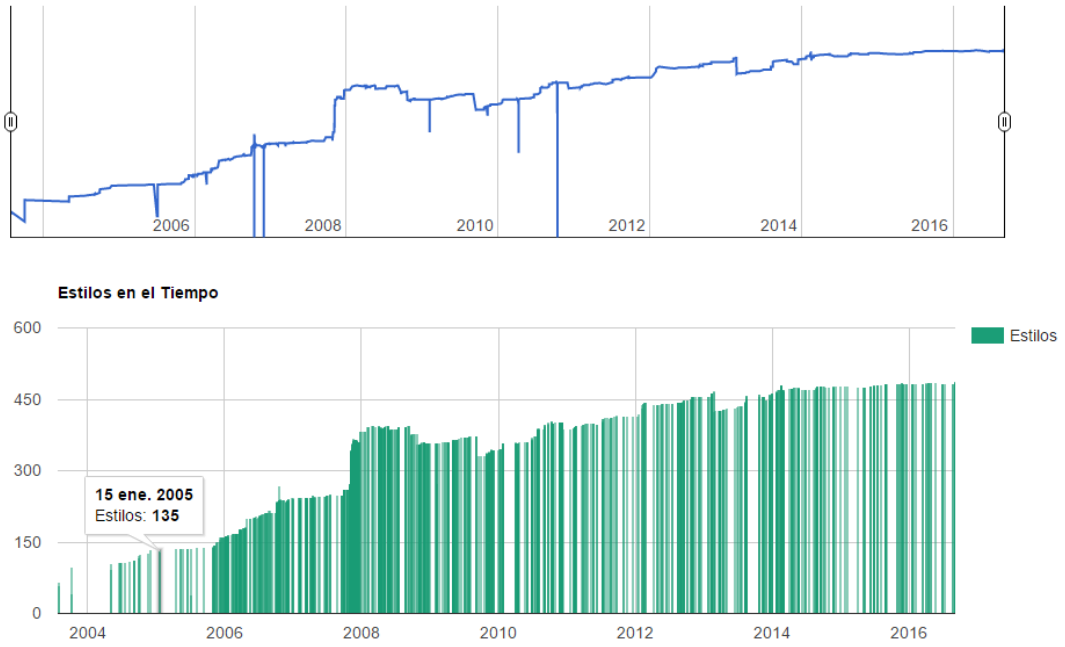


Figura 5.13: Estilos en el tiempo de la página Julio Cortázar. Estilos Acumulados.

particulares durante el análisis.

Luego se obtuvo para cada artículo un gráfico en el cual se visualiza la cantidad total de estilos por revisión. El gráfico esta generado con la métrica *StylesOfRevisionsPerDay*, que a su ves utiliza la métrica *StylesInText*, aplicada a la información de markup extraída de el texto de cada revisión de los artículos. Estos gráficos nos permiten obtener información de los cambios de estilo a lo largo de la evoluciona de un articulo de forma general.

En la Figura 5.13 podemos observar como evoluciona la cantidad total de estilos aplicados en las revisiones del articulo de Julio Cortázar. En el gráfico podemos observar que uno de los mayores crecimientos se detectan a fines del 2007 y también observamos 3 caídas totales de la cantidad de estilos aplicados durante fines del 2006 y fines del 2010.

En la Figura 5.14 podemos observar como evoluciona la cantidad total de estilos aplicados en las revisiones del articulo del Papa. En el gráfico podemos observar que si bien el gráfico comienza con una cantidad de estilos aplicados equivalente a la cantidad actual, a mediados del 2002 se produce una caída abrupta de estilos a partir de la cual comenzó un crecimiento que se estabiliza en el año 2014. También observamos un crecimiento aislado muy pronunciado de la cantidad total de estilos aplicados a mediados del 2007. Ademas desde el 2005 comienzan a detectarse periodos en los que se producen

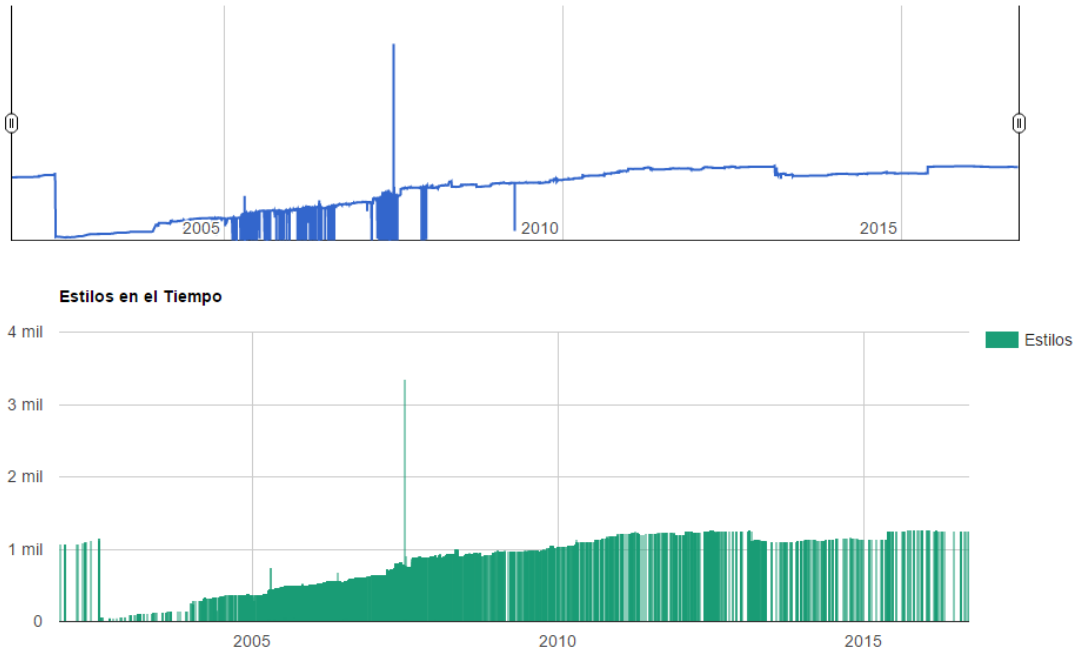


Figura 5.14: Estilos en el tiempo de la página Pope. Estilos Acumulados

repetidas caídas abruptas del total de estilos aplicados al articulo.

En la Figura 5.15 podemos observar como evoluciona la cantidad total de estilos aplicados en las revisiones del articulo de Johnny Depp. En el gráfico podemos observar múltiples caídas abruptas del total de estilos aplicados al articulo, algunas en formas aisladas y otras de forma periódica a lo largo de la historia del articulo.

En la Figura 5.16 podemos observar como evoluciona la cantidad total de estilos aplicados en las revisiones del articulo de Barack Obama. En el gráfico podemos observar múltiples caídas abruptas del total de estilos aplicados al articulo, algunas en formas aisladas y otras de forma periódica a lo largo de la historia del articulo. Particularmente este articulo es el que mas cantidad de caídas totales de estilos presenta. También se observan periodos de crecimiento, como en el año 2007, y decrecimiento, como a finales del 2008, muy abruptos junto con gran cantidad de picos aislados que marcan incrementos sustanciales de una revisión a la siguiente.

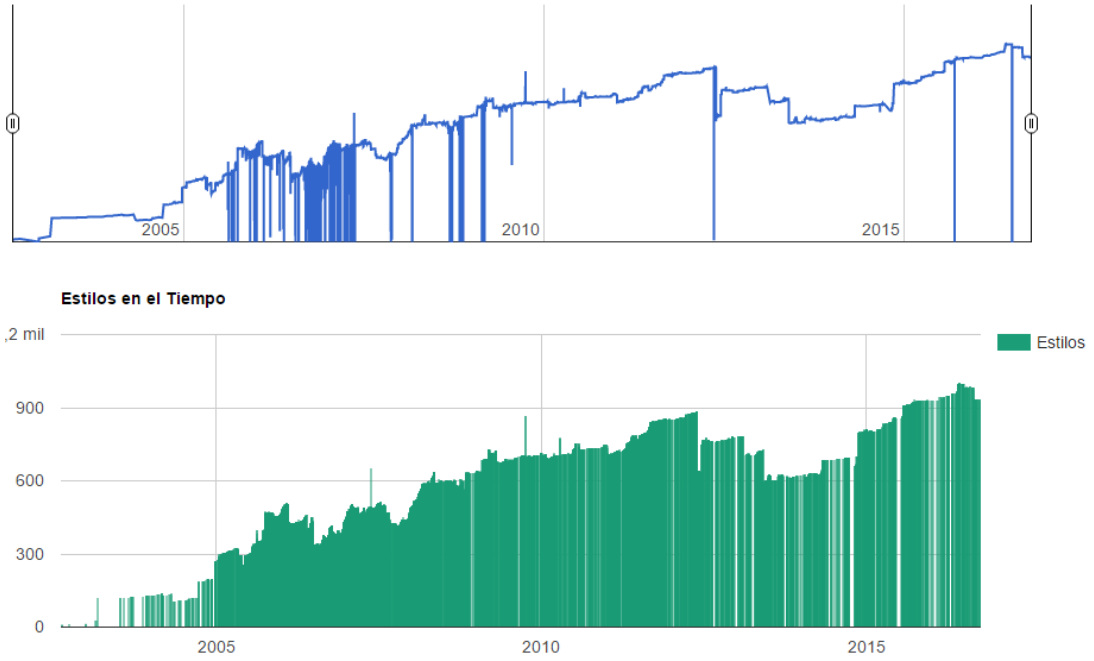


Figura 5.15: Estilos en el tiempo de la página Johnny Depp. Estilos Acumulados.

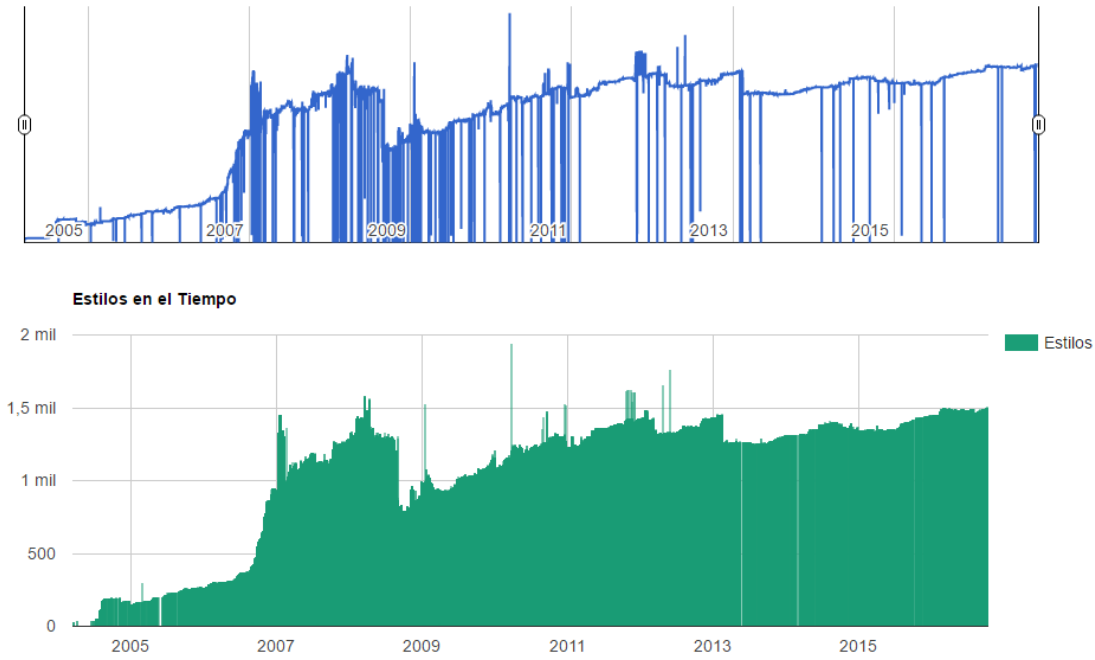


Figura 5.16: Estilos en el tiempo de la página Barack Obama. Estilos Acumulados.

5.5. Análisis de resultados

Dado que en los resultados obtenidos anteriormente hemos obtenido respuesta a los siguientes puntos mencionados al inicio de esta parte del trabajo:

- Saber cuál fue el crecimiento o decrecimiento en cada revisión respecto a su tamaño.
- Conocer qué autores y en qué medida contribuyeron a la página.
- Medir el nivel de actividad en la edición de una página en días.
- Poder analizar el cambio de estilos entre dos revisiones.
- Obtener información de los cambios de los estilos para una página a lo largo de su historial de revisiones, tanto de forma general como específica por cada tipo de estilo.

Luego de analizar todos los resultados anteriores en busca de hitos o sucesos de interés los contengan relación con sucesos históricos o circunstancias de interés en las ediciones realizadas en Wikipedia, se encontraron varios hechos relevantes de los cuales algunos se detallaran a continuación.

5.5.1. Barack Obama

El 5 de noviembre del 2008 el gráfico de la Figura 5.17 se detectó un incremento abrupto de la actividad de los editores en el artículo, el más alto en la historia del mismo. También se detectó en el gráfico de la Figura 5.18 un incremento del tamaño del artículo el cual marca el comienzo de una nueva etapa de crecimiento del mismo. Por último en el gráfico de la Figura 5.19, el cual contiene la suma de los estilos, se detectó un incremento de los estilos aplicados al artículo acompañando de igual forma el crecimiento del artículo.

Esto se debió a que muchos editores comenzaron a la vez a realizar múltiples ediciones agregando y eliminando contenido en busca de un balance frente a un crecimiento abrupto del artículo. Anteriormente a que se detecte este incremento de actividad ya se había detectado un periodo de crecimiento del articulo tanto en el nivel de actividad como en el tamaño y la cantidad de estilos aplicados. Este periodo de crecimiento coincidió con el periodo electoral de los EEUU por lo cual se cree que el nivel de actividad detectado el 5 de noviembre del 2008 puede estar relacionado con la victoria de Obama en las elecciones presidenciales que lo convirtieron en presidente durante su primer mandato las cuales fueron el 4 de noviembre del 2008.

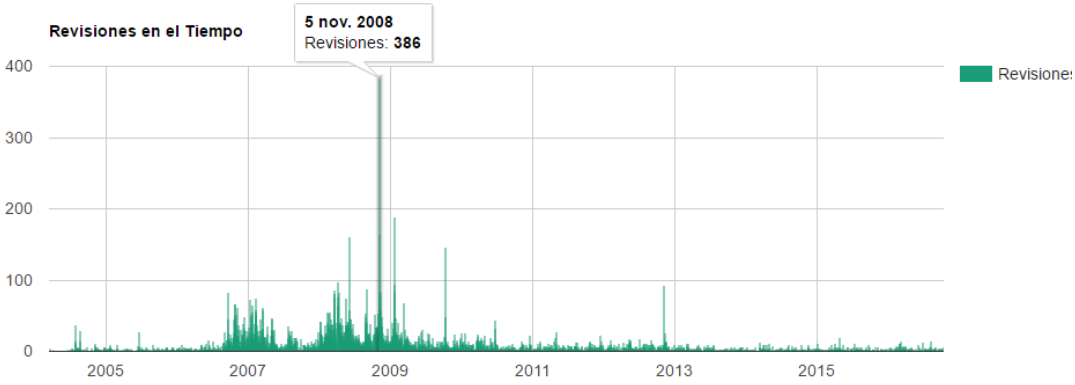


Figura 5.17: Revisiones en el tiempo de la página Barack Obama. 5 de noviembre del 2008.

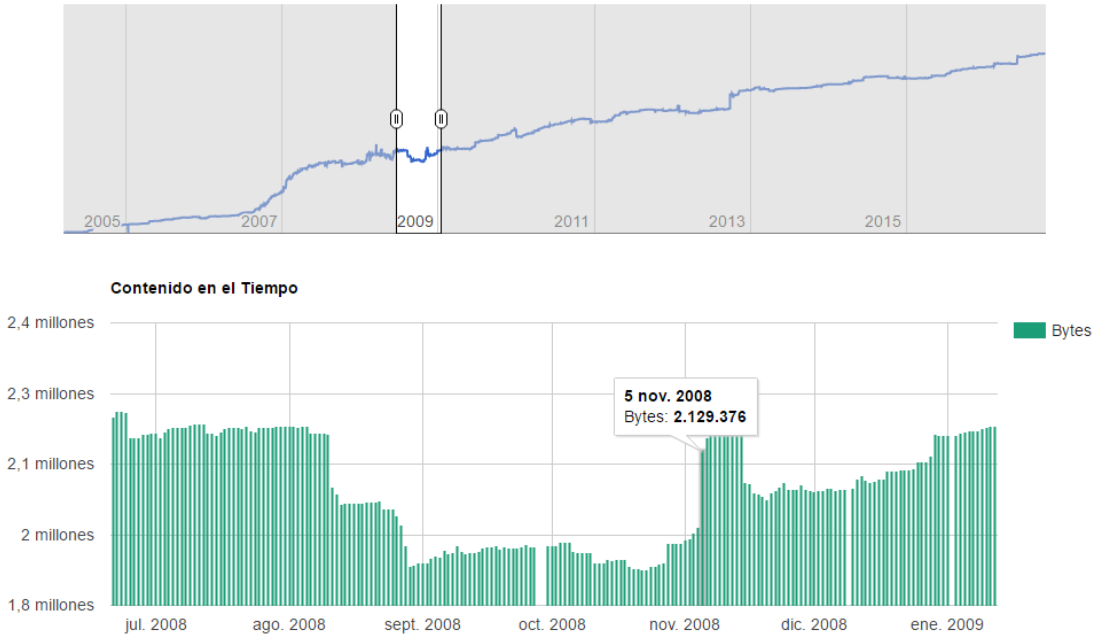


Figura 5.18: Contenido en el tiempo de la página Barack Obama. 5 de noviembre del 2008.

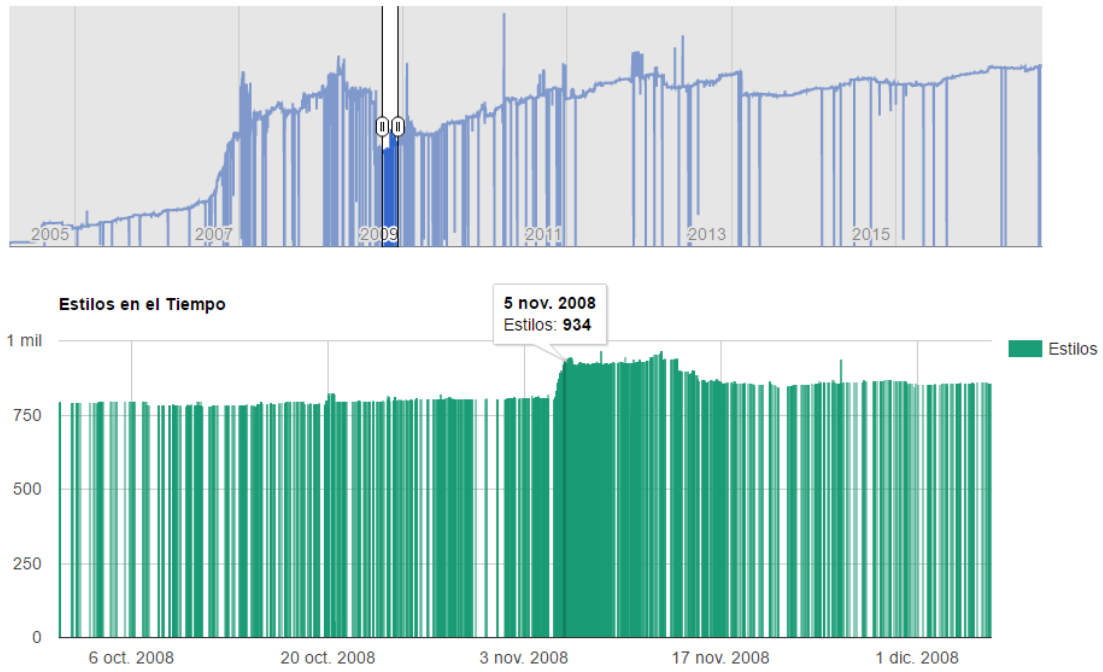


Figura 5.19: Estilos en el tiempo de la página Barack Obama. 5 de noviembre del 2008.

5.5.2. Pope - Elecciones Papales

En este caso como podemos observar en el gráfico de la Figura 5.20 las dos fechas en las que el artículo del papa tuvo sus mayores niveles de actividad coincide con las fechas en las que fueron electos los papas Benedicto XVI, 19 de abril del 2005, y Francisco, 13 de marzo del 2013.

Además se observa para cada una de esas fechas cambios relevantes tanto en contenido o en estructura como puede observarse en los gráficos 5.21 y 5.22 los cuales se corresponden con las fechas de la elección de los papas Benedicto XVI y Francisco respectivamente.

Por último es importante aclarar que la página del Papa(Pope) es una pagina referente al cargo por lo cual el contenido de la misma varía según quien lo posea y es por ello que estos eventos desencadenaron una tendencia hacia la edición de este artículo.

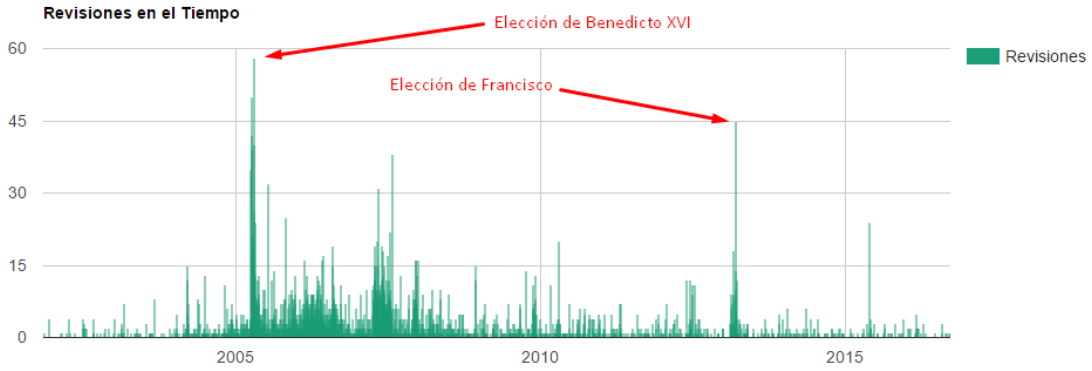


Figura 5.20: Revisiones en el tiempo de la página Papa (Pope).Elecciones Papales

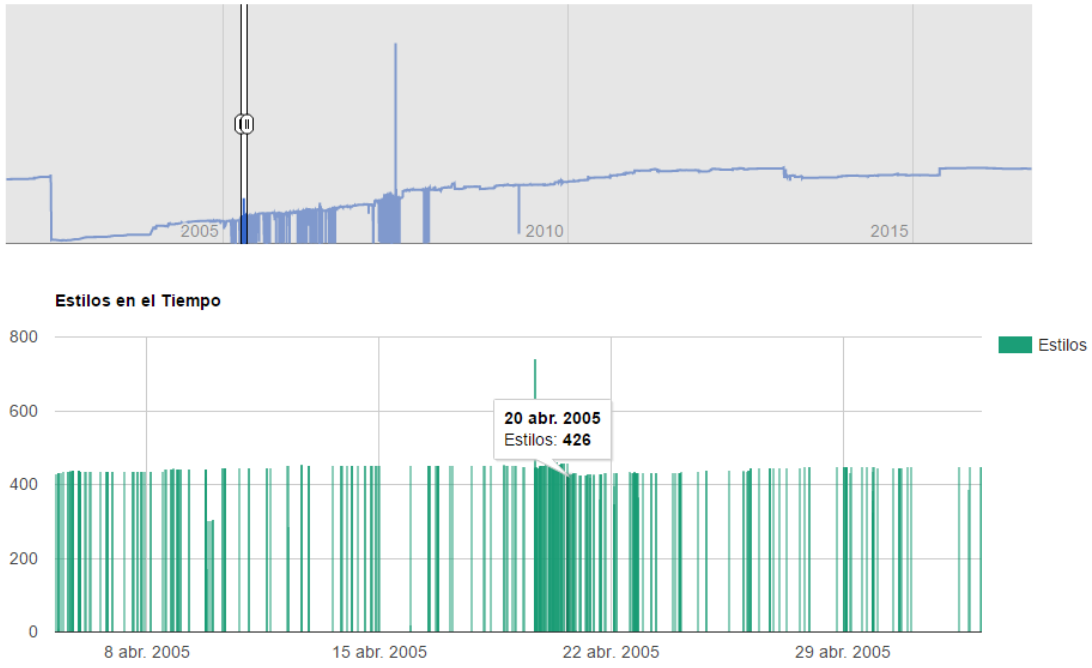


Figura 5.21: Estilos en el tiempo de la página Papa (Pope). 20 de abril del 2005.

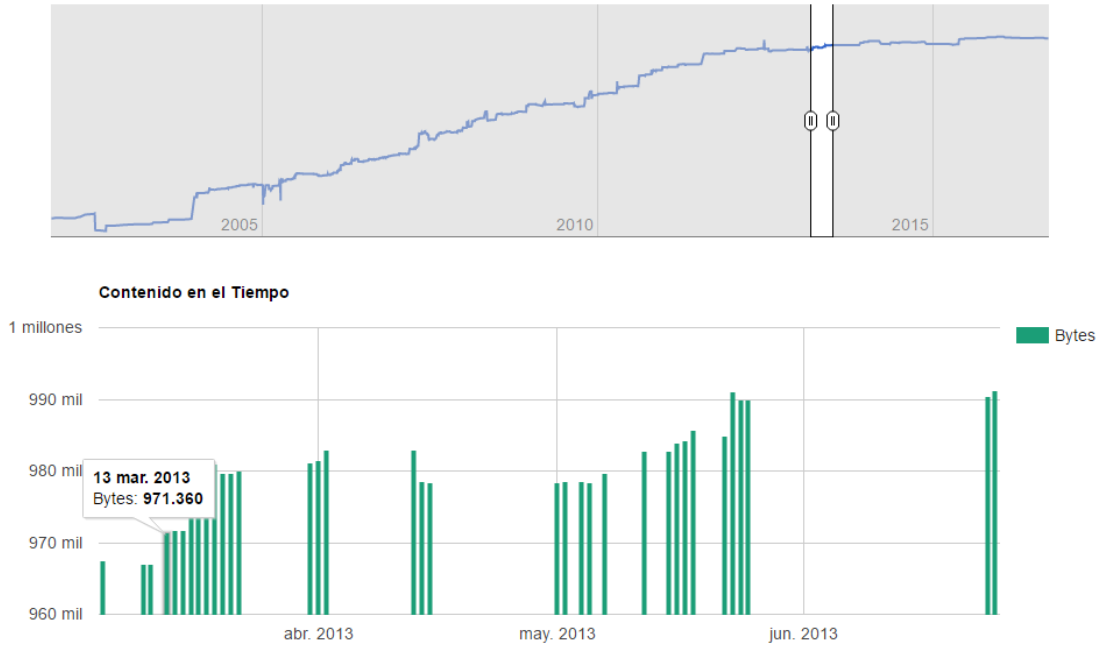


Figura 5.22: Contenido en el tiempo de la página Papa (Pope). 13 de Marzo del 2013.

5.5.3. Johnny Depp y Pope - Periodos de vandalismo

Referente a circunstancias de interés en las revisiones de los artículos de una wiki presentamos una relación entre nuestro trabajo y la detección automatizada de vandalismo.

Vandalismo en lo referente a ediciones en una wiki es toda revisión maliciosa o con intención de dañar el contenido del artículo. Esta se puede dar en forma de eliminaciones de contenido, agregando contenido contraproducente o a partir de la malversación del contenido existente en un artículo.

Además podemos observar en los gráficos de las Figuras 5.23 y 5.24, que representan la suma de estilos presente para cada una de las revisiones, periodos en los que se detectan múltiples caídas abruptas o picos aislados en la cantidad de estilos aplicados a los artículos. Por lo cual partiendo de la definición de vandalismo anterior y de la comprobación de las revisiones en Wikipedia, en las cuales se observan casos en que el contenido es completamente eliminado o remplazado por contenido sin sentido, podemos afirmar que esos fueron hechos de vandalismo dando lugar a periodos de vandalismo. Y estos periodos suelen finalizar al aplicarse medias de seguridad a los artículos que previenen los actos te vandalismo.

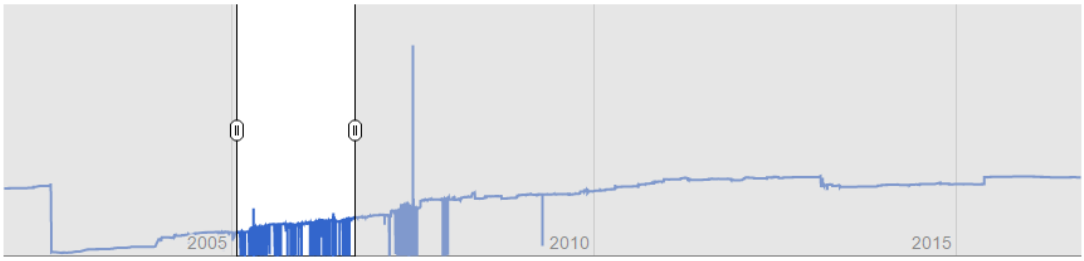


Figura 5.23: Estilos en el tiempo de la página Papa (Pope). Febrero de 2005 hasta Agosto del 2006.

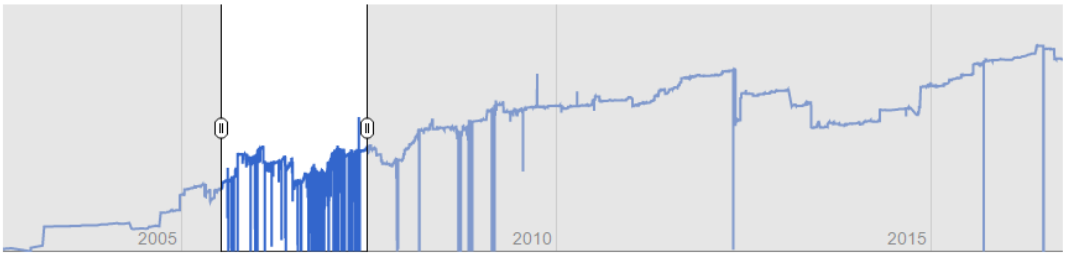


Figura 5.24: Estilos en el tiempo de la página Johnny Depp. Julio del 2005 a Mayo del 2007.

También podemos observar que luego de los periodos de vandalismo aun ocurren casos en los que se detectan caídas totales de estilos o estructura. De esta observación por lo tanto consideramos que el análisis estructural o de estilos puede ser una variable de importancia para la detección de hechos de vandalismo en artículos

Parte II

Análisis de datos

Introducción

En esta Parte del trabajo se habla de como continuar analizando información generada por el prototipo explicado en el Capitulo 3. Con este fin se explican herramienta para el manejo de datos y calculo estadístico y técnicas de aprendizaje no supervisado.

Las técnicas que se utilizan en esta Parte del trabajo son Cluster Analysis[análisis de grupos] utilizando el algoritmo K-means y minado de reglas de asociación utilizando los algoritmos Eclat y Apriori. Luego se detallara el conjunto de datos que conforma la información a analizar obtenidos de prototipo.

Se presentaran también diferentes estudios en los que con las técnicas y herramientas se buscan obtener resultados puntuales relacionados con la búsqueda de patrones dentro del conjunto de datos. Por cada estudio se detallaran objetivos del mismo, una descripción de la forma de evaluación y un análisis de los resultados.

Finalmente se presentaran las conclusiones de este trabajo y los posibles trabajos futuros.

Capítulo 6

Herramientas de análisis

6.1. Introducción

En este Capitulo se presentaran las herramientas con las que se estudiará la información obtenida mediante el prototipo descrito en el Capitulo 4. Para esto se describirá el entorno y lenguaje de desarrollo R especializado en cálculo estadístico.

Luego se hablara sobre las técnicas de aprendizaje no supervisado mas exactamente sobre Cluster Analysis[análisis de grupos] y Minado de reglas de asociación.

Por ultimo se presentara el conjunto de datos a utilizar para el análisis

6.2. R para computación estadística

R es un ambiente de desarrollo que provee un lenguaje para la manipulación de datos y la realización de cálculos estadísticos. Inicialmente fue creado por Ross Ihaka y Robert Gentleman en el departamento de estadística de la universidad de Auckland, Nueva Zelanda(22). El diseño de R está basado principalmente en S de Becker, Chambers and Wilks y en Scheme de Sussman (23). Es un lenguaje interpretado que permite desde programación modular a partir de funciones hasta la posibilidad de generación de hilos de procesamiento. La mayoría de la funciones en R están escritas en R, pero se pueden encontrar funciones escritas en C, C++ o FORTRAN por la eficiencia requerida en las mismas.

R posee las siguientes características según Dan Toomey(24):

- Provee una sintaxis sencilla para la operación de datos.

- Tiene un conjunto de herramientas que permiten la carga de los datos desde diversos formatos.
- Permite gestionar conjuntos de datos en memoria.
- Un gran conjunto de herramientas para análisis de datos integrada y Open Source.
- Facilidades para generación y almacenamiento de gráficos en diversos formatos.

6.3. Aprendizaje no supervisado

El aprendizaje no supervisado es una forma de aprendizaje automático que permite descubrir relaciones desconocidas en los datos. Es la contra partida del aprendizaje supervisado el cual busca predecir resultados en base a relaciones ya conocidas. Por esto se dice que las técnicas de análisis no supervisado no suelen terminar en sí mismos sino que son una forma de encontrar relaciones y patrones que puedan ser usados para posteriormente generar modelos predictivos. Por ejemplo Zumel et. al.(15) en su trabajo los menciona como procesos exploratorios por esto mismo.

En este trabajo utilizaremos dos tipos de métodos no supervisados. Cluster analysis o análisis de grupos, este método nos permite encontrar grupos en nuestros datos con características similares. Association rule mining o minería de reglas de asociación, este método nos permite encontrar elementos o propiedades en nuestros datos que tienden a ocurrir en conjunto.

6.3.1. Análisis de grupos

En el análisis de grupos (25) se busca agrupar los datos en clusters, donde cada dato perteneciente a un cluster sea más similar a otro dato en el mismo cluster que a otro dato en otro cluster. El análisis de grupo está fuertemente ligado con el problema de estimación de densidad, si uno piensa su información en un gran espacio de varias dimensiones puede entonces querer encontrar regiones de ese espacio donde hay mayor densidad de datos. Si esas regiones son distintas entonces obtenemos clusters.

Para medir estos agrupamientos se usan conceptos de similitud y disimilitud, en los cuales la disimilitud puede ser considerada distancia por lo cual los puntos en un cluster suelen estar más cerca uno de otros que con puntos en otros clusters. Existen diferentes conceptos de distancia entre los que se encuentran:

- Distancia Euclidiana.
- Distancia de Hamming.

- Distancia Manhattan.
- Similitud Coseno.

En este trabajo se utilizará el algoritmo K-means. Este algoritmo de agrupamiento trabaja con datos numéricos permitiendo utilizar diversos tipos de distancias para el calculo, en este trabajo se utilizara la distancia Euclidiana la cual es conocida como la distancia ordinaria y se deduce del teorema de Pitagoras (26). Su mayor desventaja radica en que se debe elegir desde el inicio el número de clusters o grupos deseados conocido como k. Por otro lado tiene la ventaja de ser fácil de implementar y es más rápido que otros algoritmos de agrupamiento sobre conjuntos de datos grandes. Aunque no es el caso de la información que se maneja en este trabajo este algoritmo suele trabajar mejor en conjuntos de datos que posean una distribución similar a la Gausiana.

El algoritmo no garantiza tener un único punto de finalización que brinde siempre el mismo resultado ya que como el algoritmo trabaja en dos, una en la que los centroides se eligen aleatoriamente y otra en la que se van recalculando en la medida que cambie el contenido de los mismos, el resultado depende del punto inicial de los centroides, cada grupo posee un centroide en su centro y los elementos pertenecen al grupo del centroide que poseen mas cerca. Ademas Por lo cual es una buena práctica correr el algoritmo muchas veces con inicios aleatorios utilizando el que tenga la menor suma total de cuadrados o WSS por su nombre en inglés “total within sum of squares”. Además para asegurarnos que los clusters encontrados son clusters reales podemos realizar iteraciones a partir de las cuales evaluar los cluster de cada iteración con los de la iteración anterior con el coeficiente de Jaccard el cual brinda una medida de similitud entre dos grupos o clusters, si se obtiene un valor menor de 0.5 se disuelve el cluster ya que probablemente no fuera un cluster real. Luego de finalizadas las iteraciones la media del coeficiente de Jaccard es considerado el nivel de estabilidad de cada cluster y se puede medir de la siguiente forma:

- Menos de 0.6 puede considerarse inestable.
- Entre 0.6 y 0.75 indica un nivel de estabilidad bajo.
- Entre 0.75 y 0.85 un nivel de estabilidad medio.
- Más de 0.85 un nivel de estabilidad alto.

A mayor estabilidad es más probable que el cluster denote un patrón dentro de nuestros datos.

También otro punto importante a tener en cuenta es la selección de la cantidad k de clusters para utilizar en los algoritmos. Para dicha selección se utilizan los criterios de Calinski-Harabasz Index (27), abreviado como “ch”, y el “average silhouette width” abreviado como “asw”(28). Para utilizar ambos criterios en conjunto la opción que se

utilizo fue realizar un gráfico de ambos valores para un rango de K tomando los valores más altos o en común de ambos criterios.

6.4. Reglas de asociación

La minería de reglas de asociación (15) es utilizado para encontrar elementos o atributos que suelen ocurrir en conjunto. La unidad mínima utilizada para el minado de reglas es llamado transacciones, las mismas están compuestas de elementos no continuos.

Dependiendo que se desee evaluar la transacciones pueden ser distintos elementos, por ejemplo en nuestro trabajo pueden ser desde una revisión con sus componentes como elementos de la transacción hasta un artículo con sus diferentes métricas como elementos de la transacción. Para obtener reglas de un conjunto de transacciones primero se buscan todos los conjuntos de elementos que ocurren más comúnmente en las transacciones y a partir de estos conjuntos de elementos se transforman en reglas.

En este trabajo se utilizaron Eclat (29) y Apriori (30) como algoritmos de minería de reglas estos poseen diferentes métricas y posibilidades de configuración.

Eclat es utilizado para el minado de elementos frecuentes en un conjunto de transacciones, para ello computa el soporte que representa la frecuencia con la que un elemento por sí mismo o un conjunto de elementos ocurren en las transacciones. Por lo tanto el soporte o "support" mide el total de ocurrencia del elemento o conjunto de elementos X en el total de las transacciones T siendo el valor del mismo $\#X/\#T$.

Además Eclat nos permite seleccionar el número máximo y mínimo de elementos en X lo cual puede servirnos para por ejemplo si nos interesa ver cuantos elementos se relacionan frecuentemente entre sí con un mínimo de 2 podemos evitar los resultados referentes a el nivel de ocurrencia de un único elemento. También nos permite seleccionar un soporte mínimo para filtrar los resultados.

Apriori por otro lado es utilizado para el minado de reglas de asociación de la forma "si X entonces Y" las cuales son evaluadas por su soporte, confidencia y elevación o por sus nombres en inglés "support", "confidence" y "lift" respectivamente. El soporte al igual que Eclat es el número de transacciones que contienen X dividido el total de transacciones. La confidencia entonces es el soporte de la unión entre X e Y sobre el soporte de X, o dicho de otra forma en cuantas ocasiones aparece Y cuando esta X. La elevación o lift es el soporte de la unión de X e Y sobre el producto del soporte de X e Y por separado, esta métrica es una medida de calidad que compara la frecuencia con que se observa un patrón con la que se esperaría ver ese patrón por casualidad.

Por lo cual si el valor es cercano a 1es mas probable que el patrón sea algo que se está observando por casualidad, pero a mayor tamaño es más probable que el patrón sea real.

En cuanto a la configuración Apriori nos permite configurar los mismos límites, limitar el tamaño en X y limitar el mínimo de soporte y confidencial para los resultados. Además nos permite limitar qué elementos nos interesa ver en los resultados y en qué lugar de la relación si como X o como Y. Estas configuraciones nos permiten un análisis más dinámico de los resultados permitiendo una mejor exploración del conjunto de datos.

6.5. Conjunto de datos

Los datos que se utilizaron en los análisis que se presentarán mas adelante se obtuvieron del prototipo descrito en el capítulo 4, en el mismo se mencionó una funcionalidad que permite exportar en formato json la información que se obtuvo del prototipo. El formato fue seleccionado dado que R nos provee facilidades para la lectura de diversos tipos de formatos entre ellos json.

Los datos en formato json se encuentran particionados en un archivo por artículo el cual contiene la siguiente información:

- Nombre del artículo.
- Distribución de aportes por autor.
- Total de revisiones.
- Fecha y hora de cada revisión.
- Estilos aplicados por cada revisión.
- Tamaño de contenido de cada revisión.
- Categorías a las que pertenece el artículo por cada revisión.
- Cantidad de revisiones día a día

A partir de esta información se generaron diversas estructuras de datos en R para la realización de los distintos análisis.

Sobre la información contenida en estas estructuras podemos decir que se cuentan con 100.000 artículos y más de 2.000.000 de revisiones. Estos artículos fueron tomados de forma aleatoria, como se describe en el Capitulo 4, de la Wikipedia en inglés.

Capítulo 7

Análisis de Resultados

7.1. Introducción

En este Capítulo utilizaremos las técnicas descritas en el Capítulo 6 para analizar la información obtenida y buscar patrones dentro del contenido de la misma.

Por ello en las siguientes secciones se detallarán los diversos estudios realizados. Cada sección contendrá información sobre:

- Nombre otorgado para identificar el caso de estudio.
- Información sobre que se desea estudiar o analizar.
- Los datos utilizados para el estudio, junto con qué métricas los generó o de qué información derivan.
- Desarrollo del caso de estudio con información sobre metodologías,algoritmos y configuraciones utilizadas.
- Presentación y análisis de los resultados obtenidos.

7.2. Estudio 1 - Cluster Analysis de la estructura del contenido de las revisiones

7.2.1. Nombre

CA_EstructuraContenidoRevisiones_1

7.2.2. Objetivo del estudio

En este estudio se desea observar qué tipo de relaciones hay entre la información que se obtuvo de todos los artículos.

7.2.3. ¿Con qué herramientas se realizaron?

Para este estudio se utilizaran los datos de las revisiones de varias de las páginas, no se utilizaron el total de las revisiones durante una única iteración debido al volumen de las mismas. En su lugar se realizaron múltiples iteraciones con valores incrementales del número de revisiones tomadas. La información contenida en dichas revisiones deriva de las siguientes métricas:

- *RevisionsPerDay*, de la cual obtenemos cuántas revisiones se realizaron en cada fecha.
- *BytesOfRevisionPerDay*, a partir del cual obtenemos los tamaños de los artículos revisión a revisión.
- *#OcurrencesOfStyle*, el cual aplicamos para obtener las ocurrencias de cada estilo para cada revisión.

Para este primer estudio se aplicó el algoritmo de clustering K-means por lo cual los valores utilizados fueron valores numéricos continuos por las limitaciones ya mencionadas de este tipo de algoritmo.

7.2.4. Metodología y configuraciones

Para realizar este estudio se aplicó el algoritmo de clustering K-means para detectar la forma en que las revisiones se agrupaban.

Por cada iteración se tomó el conjunto de datos y se lo escalo, modificando cada una de las variables para posean una media igual a 0 y una desviación estándar igual a 1, para solucionar problemas respecto a los valores de las distintas variables. Luego se debió obtener el número de cluster óptimo para aplicar el mismo, esto se realizó utilizando dos criterios:

- Criterio CH o índice de Calinski-Harabasz.
- Criterio asw o average silhouette width.

Numero de Articulos	CH	ASW
3	17	11
4	18	18
5	20	20

Tabla 7.1: Recomendaciones de clusters.

Posteriormente se aplicó un algoritmo de K-means con los diferentes números de clusters recomendados buscando la configuración con mayor estabilidad.

7.2.5. Resultados y análisis

A partir de las pruebas realizadas se pudo comprobar que el número de clusters recomendados y con mayor estabilidad incrementa al aumentar el número de artículos presentes en la prueba. Esto puede observarse en la Tabla 7.1 en la que la primera columna indica el numero de artículos utilizados, las siguientes indican el numero de clusters recomendados por los criterios Calinski-Harabasz Index, abreviado como "CH", y el "average silhouette width" abreviado como "asw". Se utilizo un numero acotado de artículos debido a la gran cantidad de clusters que se requerían en la medida que se agregaban artículos. Este resultado se debe a que las revisiones de un mismo articulo tienden a agruparse, por lo cual se generan uno o más clusters dedicados solamente a un artículo ya que la diferencias estructural entre las revisiones del mismo suelen ser pequeñas.

Por lo cual se observa que no se puede determinar un significado de interés a dichos agrupamientos.

7.3. Estudio 2 - Cluster Analysis de la evolución en la estructura del contenido de las revisiones

7.3.1. Nombre

CA_RevisionesEvolEC_2

7.3.2. Objetivo del estudio

En este estudio se busca encontrar relación en la forma en que los artículos evolucionan o dicho de otra forma relación en la forma de edición de los editores.

7.3.3. ¿Con qué herramientas se realizaron?

A partir de los resultados de *CA_EstructuraContenidoRevisiones_1* se decidió modificar la información a fin de representar la evolución estructural y del contenido de la misma en lugar de el contenido y estructura exacta de la revisión. Por lo cual la información utilizada para este estudio es derivada de la utilizada para el estudio *CA_EstructuraContenidoRevisiones_1*. Además se utilizará información de la suma de los estilos, y sobre como cambia el numero de categorías en la que se encuentra un artículo por cada revisión.

Además se utilizara el mismo algoritmo de clustering, K-means.

7.3.4. Metodología y configuraciones

Primero se generó un nuevo conjunto de información el cual contenía las diferencias entre las revisiones respecto a su estructura y su contenido. Para generar este conjunto de información se tomo el conjunto de información utilizado en *CA_EstructuraContenidoRevisiones_1*, el cual contenía información de la estructura y contenido al momento de cada revisión, y se procedió a por cada articulo computar la diferencia entre una revisión y su revisión anterior a fin de obtener así los cambios entre revisiones de cada articulo y con dichos cambios representar la evolución de los artículos. A este nuevo conjunto de información se le aplicaron los algoritmos para detectar el número óptimo de clusters a utilizar.
Luego se realizaron iteraciones con los valores recomendados para k para encontrar el conjunto de clusters más estables utilizando los coeficientes de Jaccard.

7.3.5. Resultados y análisis

En primera instancia el algoritmo para determinar el número de clusters recomendable sugirió dos configuraciones una de 3 y otra de 5 clusters. Para verificar que configuración era mejor se realizaron múltiples iteraciones para determinar los coeficientes de Jaccard y calcular los niveles de estabilidad de cada cluster.

De estos resultados concluimos que el mejor número de clusters a utilizar eran 3 y nos dedicamos al estudio de los mismo. Para ello se procedió a estudiar sus características como valores máximos,mínimos, medias y medianas.A Partir de esto observamos

Cluster 1	Cluster 2	Cluster 3
0.5840954	0.6344311	0.9976429

Tabla 7.2: Estabilidad en 3 clusters.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
0.4816930	0.7260933	0.7998937	0.9981883	0.2649629

Tabla 7.3: Estabilidad en 5 clusters.

que estos cluster representaban 3 tipos de ediciones:

- **Cluster 1 - Ediciones vandálicas**: Son ediciones en las que podemos observar caídas abruptas en la estructura o contenido del artículo.
- **Cluster 2 - Ediciones de Corrección**: Son ediciones en las que se revierten los cambios realizados en las revisiones vandálicas, suelen estar compuestas de valores exactamente opuestos a dichas revisiones.
- **Cluster 3 - Ediciones Comunes**: son las ediciones que se mantienen dentro de un margen normal de modificación dentro del artículo.

7.4. Estudio 3 - Minería de reglas de asociación de estilos

7.4.1. Nombre

PA_Estilos_5

7.4.2. Objetivo del estudio

En este estudio se quiere ver si existe co ocurrencia de cambios en los estilos, en este experimento solo se busca identificar si existen patrones ligados a cambios de agregado o eliminación de estilos.

Por ello en este estudio se buscará encontrar patrones del tipo X entonces Y, como por ejemplo si se agregan o eliminan elementos del estilo X se agregan o eliminan elementos del estilo Y.

7.4.3. ¿Con qué herramientas se realizaron?

Para este estudio se utilizaran del conjunto de datos generados en *CA_RevisionesEvolEC_2* la evolución de los distintos estilos en cada revisión.

Además para la detección de patrones se utilizaran métodos de minado de reglas de asociación por lo cual como se explicó en el Capítulo 6 se requerirá modificar este conjunto de datos para que sea compatible con dichos métodos.

7.4.4. Metodología y configuraciones

Para este estudio en primera instancia se generó un nuevo conjunto de datos a partir del que contenía la información de evolución de los estilos de todos los artículos. Para esto esto se factorizo el contenido de la siguiente forma:

- En los estilos que se detectaba una disminución de su aplicación en la revisión se utilizó el valor *"-Eliminado"*.
- En los estilos que se detectaba un incremento de su aplicación en la revisión se utilizó el valor *"+Agregado"*.
- En los estilos en los que no se detectaba cambios fueron eliminados de la revisión para evitar así pérdida de información valiosa de cambios

Luego de la factorización del conjunto de datos al mismo se le aplicó una transformación a un tipo de dato conocido como transacciones que es con el que trabajan los algoritmos de reglas de asociación.

Los algoritmos que se evaluaron fueron Eclat y Apriori. Eclat se configuró con un soporte mínimo de 0.01 y un tamaño mínimo de elementos de 2 para omitir los estilos que se encontraran de forma individual. Apriori también se configuró con un mínimo de 2 elementos para omitir elementos vacíos en el valor de X y con un soporte mínimo de 0.01 al igual que la confidencia mínima seleccionada.

7.4.5. Resultados y análisis

Por parte de la evaluación realizada con Eclat se obtuvieron 20 elementos que se presentaran a continuación en la Tabla 7.4. Esta tabla esta conformada por 3 columnas, la primera columna con un Identificador del resultado, segunda por los elementos que co ocurren y la frecuencia con la que co ocurren, y la tercera con el soporte obtenido para cada elemento. Particularmente la forma correcta de leer la tabla, tomando de ejemplo

ID	Estilos	Soporte
[1]	{blockquote=+Agregado,includeonly=+Agregado}	0.03391964
[2]	{internal=+Agregado,indent1=+Agregado}	0.02732596
[3]	{internal=+Agregado,bulletedelement=+Agregado}	0.02166337
[4]	{internal=-Eliminado,indent1=-Eliminado}	0.01552891
[5]	{internal=-Eliminado,bulletedelement=-Eliminado}	0.01432017
[6]	{blockquote=+Agregado,big=+Agregado}	0.01347192
[7]	{external=+Agregado,indent1=+Agregado}	0.01335142
[8]	{bulletedelement=+Agregado,indent1=+Agregado}	0.01320869
[9]	{heading2=+Agregado,includeonly=+Agregado}	0.01307539
[10]	{reference=+Agregado,indent1=+Agregado}	0.01290868
[11]	{blockquote=-Eliminado,includeonly=-Eliminado}	0.01251420
[12]	{reference=+Agregado,bulletedelement=+Agregado}	0.01225568
[13]	{heading3=+Agregado,bulletedelement=+Agregado}	0.01203866
[14]	{external=+Agregado,bulletedelement=+Agregado}	0.01177999
[15]	{heading2=+Agregado,bulletedelement=+Agregado}	0.01153177
[16]	{italic=+Agregado,internal=+Agregado}	0.01138919
[17]	{blockquote=+Agregado,bulletedelement=+Agregado}	0.01124470
[18]	{s=+Agregado,includeonly=+Agregado}	0.01112287
[19]	{external=-Eliminado,indent1=-Eliminado}	0.01038630
[20]	{italic=+Agregado,bulletedelement=+Agregado}	0.01021900

Tabla 7.4: Estudio 3 resultados de Eclat.

el elemento de identificador 2, es "La aplicación del elemento internal, que representa a los enlaces internos, co ocurre con la aplicación del elemento indent1, que representa a la sangría de primer nivel, con un soporte de 0.033".

Esto nos permite saber cuales son los 20 elementos que suelen co ocurrir con mayor frecuencia en el conjunto de datos y bajo qué evento se encuentran. Por lo cual a partir de esta información podemos realizar afirmaciones del siguiente tipo:

- El estilo de enlace interno y la sangría de tipo 1 suelen ser insertadas durante la misma revisión en el 2,7 % de las revisiones del conjunto de datos. Visible en el resultado con identificador 2.

- También el estilo de enlace interno es insertado en la misma revisión con el estilo enumerativo en el 2,1 % de los casos analizados. Visible en el resultado con identificador 3.
- Ambas reglas poseen su contra parte de eliminación conjunta. Visible en los resultados con identificadores 4 y 5 respectivamente.

Por otro lado la ejecución de Apriori nos brindo las mismas 20 reglas dado el que el soporte mínimo requerido era el mismo y todas superan el coeficiente mínimo requerido. A continuación en la Tabla 7.5 se presentan dichas reglas pero en este caso ordenado por la confidencia que es un tipo de información que nos brinda Apriori. Esta tabla cuenta con 5 columnas, una con el identificador del resultado, otra con los elementos que componen el resultado y 3 columnas indicando el grado de soporte,confidencia y elevación del resultado. Los elementos de la tabla se leen de la forma X entonces Y $[X => Y]$ refiriéndose a que cuando ocurre X también lo hace Y con un cierto grado de soporte, confidencia y elevación.

La información que nos brinda Apriori extiende a la que nos brindo Eclat con información como la confidencia y la elevación. Esto nos permite saber en qué medida se cumple una regla dentro del conjunto de datos, confianza, o cuán real puede llegar a ser una regla, elevación. A partir de esto podemos realizar afirmaciones del tipo:

- El 60 % de las veces que se eliminan enlaces internos también se eliminan sangrías de tipo 1 además esto ocurre con una frecuencia 12 veces mayor a la que sería si fuera un suceso aleatorio. Esto se puede observar en el resultado con identificador 2.
- También podemos observar que en el 45 % de los casos que se insertan elementos listados también se insertan enlaces externos. También una frecuencia 9 mayor a la aleatoria. Esto se puede observar en el resultado con identificador 8.

De estos resultado podemos notar muchas más relaciones con un alto porcentaje de confidencia y elevación de lo cual podemos afirmar que hay un grado de relación entre cómo se aplican y eliminan estilos en las revisiones. También observamos que el soporte limita principalmente las reglas a los estilos que aparecen frecuentemente en las revisiones por lo cual disminuyendo el soporte podríamos observar mayor cantidad de casos en los que haya relaciones fuertes entre los estilos. Por último podemos observar que en estas relaciones no hay relaciones del tipo si agrego X entonces elimino Y o viceversa.

Id	X =>Y	Soporte	Confidencia	Elevación
[1]	{s=+Agregado}=>{includeonly=+Agregado}	0.01112287	0.7554439	13.215.852
[2]	{external=-Eliminado}=>{indent1=-Eliminado}	0.01038630	0.6004355	20.110.261
[3]	{includeonly=+Agregado}=>{blockquote=+Agregado}	0.03391964	0.5933955	7.576.177
[4]	{indent1=-Eliminado}=>{internal=-Eliminado}	0.01552891	0.5201064	12.117.157
[5]	{includeonly=-Eliminado}=>{blockquote=-Eliminado}	0.01251420	0.5188473	16.603.750
[6]	{external=+Agregado}=>{indent1=+Agregado}	0.01335142	0.5156500	9.520.728
[7]	{indent1=+Agregado}=>{internal=+Agregado}	0.02732596	0.5045341	6.857.714
[8]	{external=+Agregado}=>{bulletedelement=+Agregado}	0.01177999	0.4549593	5.379.638
[9]	{heading3=+Agregado}=>{bulletedelement=+Agregado}	0.01203866	0.4488496	5.307.395
[10]	{italic=+Agregado}=>{internal=+Agregado}	0.01138919	0.4446570	6.043.854
[11]	{heading2=+Agregado}=>{includeonly=+Agregado}	0.01307539	0.4247463	7.430.578
[12]	{big=+Agregado}=>{blockquote=+Agregado}	0.01347192	0.4134381	5.278.571
[13]	{italic=+Agregado}=>{bulletedelement=+Agregado}	0.01021900	0.3989706	4.717.604
[14]	{heading2=+Agregado}=>{bulletedelement=+Agregado}	0.01153177	0.3746027	4.429.467
[15]	{reference=+Agregado}=>{indent1=+Agregado}	0.01290868	0.3668845	6.773.989
[16]	{reference=+Agregado}=>{bulletedelement=+Agregado}	0.01225568	0.3483254	4.118.752
[17]	{internal=-Eliminado}=>{bulletedelement=-Eliminado}	0.01432017	0.3336236	7.942.472
[18]	{internal=+Agregado}=>{bulletedelement=+Agregado}	0.02166337	0.2944522	3.481.733
[19]	{indent1=+Agregado}=>{bulletedelement=+Agregado}	0.01320869	0.2438793	2.883.736
[20]	{blockquote=+Agregado}=>{bulletedelement=+Agregado}	0.01124470	0.1435667	1.697.596

Tabla 7.5: Estudio 3 resultados de Apriori.

7.5. Estudio 4 - Minería de reglas de asociación de estilos y frecuencias

7.5.1. Nombre

PA_EstilosNivelados_6

7.5.2. Objetivo del estudio

En este estudio buscamos saber en qué medida existen patrones de estilos siendo agregados o eliminados en conjunto. Los patrones que se buscan son del tipo si se agrega o elimina X en una medida entonces se agrega o elimina Y en una medida.

7.5.3. ¿Con qué herramientas se realizaron?

Para este estudio se utilizara el mismo conjunto de datos que en el estudio *PA_Estilos_5* pero se modificara la forma en que se realizaron la factorización de los mismos.

7.5.4. Metodología y configuraciones

Para este estudio en primera instancia se generó un nuevo conjunto de datos a partir del que contenia la informacion de evolución de los estilos de todos los artículos. Para esto esto se factorizo el contenido de la siguiente forma:

- En los estilos que se detectaba una disminución de su aplicación en la revisión se utilizó los valores *"-Bajo-","-Medio-"* y *"-Alto-"* según cuantas aplicaciones del estilo se eliminaron con rangos entre (0 : -2], (-2 : -10] y (-10:-Inf).
- En los estilos que se detectaba un incremento de su aplicación en la revisión se utilizó los valores *"+Bajo+","+Medio+"* y *"+Alto+"* según cuantas aplicaciones del estilo se eliminaron con rangos entre (0 : 2], (2 : 10] y (10:Inf).
- En los estilos en los que no se detectaba cambios fueron eliminados de la revisión para evitar así pérdida de información valiosa de cambios

Luego de la factorización del conjunto de datos al mismo se le aplicó una transformación a un tipo de dato conocido como transacciones que es con el que trabajan los algoritmos de reglas de asociación.

Los algoritmos que se evaluaron fueron Eclat y Apriori. Eclat se configuró con un soporte mínimo de 0.01 y de 0.001, se limitó el tamaño mínimo de elementos a 2 para omitir los estilos que se encontraran de forma individual. Apriori también se configuró con un mínimo de 2 elementos para omitir elementos vacíos en el valor de X. Además se le configuraron soporte y confidencia mínimas primero de 0.01 y luego de 0.001

7.5.5. Resultados y análisis

Los resultados de la evaluación con Eclat fueron solamente 3 elementos, dado que los estilos se dividieron en más subelementos según el nivel de cambio es lógico que haya un menor soporte para cada uno de los elementos por lo cual se realizó el estudio nuevamente pero en esta ocasión con un valor de soporte mínimo de 0.001.

En esta segunda iteración se encontraron 460 elementos de los cuales se mostraran los primeros 30 a continuación en la Tabla 7.6.Esta tabla esta conformada por 3 columnas, la primera columna con un Identificador del resultado, segunda por los elementos que co ocurren y la frecuencia con la que co ocurren, y la tercera con el soporte obtenido para cada elemento. Particularmente la forma correcta de leer la tabla, tomando de ejemplo el elemento de identificador 2, es "La aplicación en un nivel bajo del elemento internal, que representa a los enlaces internos, co ocurre con la aplicación en un nivel bajo del elemento indent1, que representa a la sangría de primer nivel, con un soporte de 0.033".

Con estos resultados podemos observar que al dividir en múltiples valores el soporte para cada uno disminuye en gran medida dentro del conjunto de datos, además por los resultados obtenidos podemos ver que la co ocurrencia de estilos respeta también en las medidas en las que suceden sobretodo en situaciones en las que las proporciones de aplicación o eliminación son chicas.

Al ejecutar el algoritmo de Apriori nuevamente conseguimos 3 reglas relacionadas con los 3 primeros elementos obtenidos en Eclat por esto se modificó el nivel mínimo de soporte y confidencia requerido a 0.001 con lo cual se obtuvieron un total de 298 reglas de las cuales 30 se presentan a continuación en la Tabla 7.7.Los elementos de la tabla se leen de la forma X entonces Y $[X => Y]$ refiriéndose a que cuando ocurre X también lo hace Y con un cierto grado de soporte, confidencia y elevación.

Los resultados de la evaluación de Apriori nos permiten observar como la división de la medida de actividad resultó en un incremento de la confidencia máxima en los resultados así como mayor nivel de elevación en los mismos. Esto junto con que las medidas de las reglas coinciden en la mayoría de los casos, en otras palabras, que las aplicaciones o las eliminaciones coinciden en las cantidades aplicadas y eliminadas nos permite suponer una posible relación directa entre la aplicación o eliminación en con-

ID	Estilos	Soporte
[1]	{blockquote=+Bajo+,includeonly=+Bajo+}	0,02150432
[2]	{internal=+Bajo+,indent1=+Bajo+}	0,01552964
[3]	{internal=+Bajo+,bulletedelement=+Bajo+}	0,01300579
[4]	{heading2=+Bajo+,includeonly=+Bajo+}	0,00800723
[5]	{external=+Bajo+,indent1=+Bajo+}	0,00736968
[6]	{blockquote=+Bajo+,big=+Bajo+}	0,00697477
[7]	{s=+Bajo+,includeonly=+Bajo+}	0,00685411
[8]	{bulletedelement=+Bajo+,indent1=+Bajo+}	0,0068472
[9]	{reference=+Bajo+,indent1=+Bajo+}	0,00669212
[10]	{external=+Bajo+,bulletedelement=+Bajo+}	0,00639504
[11]	{reference=+Bajo+,bulletedelement=+Bajo+}	0,00583416
[12]	{blockquote=+Bajo+,cite=+Bajo+}	0,0054001
[13]	{heading3=+Bajo+,bulletedelement=+Bajo+}	0,005372
[14]	{big=+Bajo+,includeonly=+Bajo+}	0,00513143
[15]	{heading2=+Bajo+,indent2=+Bajo+}	0,00507654
[16]	{internal=-Medio-,bulletedelement=-Bajo-}	0,00461527
[17]	{heading4=+Bajo+,italicblod=+Bajo+}	0,0045879
[18]	{blockquote=-Bajo-,includeonly=-Bajo-}	0,00447196
[19]	{sub=+Bajo+,big=+Bajo+}	0,00434556
[20]	{heading2=+Bajo+,bulletedelement=+Bajo+}	0,00429922
[21]	{internal=-Alto-,indent1=-Alto-}	0,00422167
[22]	{external=+Bajo+,bulletedelement=+Bajo+,indent1=+Bajo+}	0,00408336
[23]	{internal=+Alto+,indent1=+Alto+}	0,00393608
[24]	{italic=+Bajo+,bulletedelement=+Bajo+}	0,00390621
[25]	{blockquote=+Bajo+,bulletedelement=+Bajo+}	0,00388797
[26]	{s=+Bajo+,big=+Bajo+}	0,00383853
[27]	{indent2=+Bajo+,includeonly=+Bajo+}	0,00379365
[28]	{heading2=+Bajo+,italicblod=+Bajo+}	0,00377688
[29]	{bulletedelement=+Bajo+,big=+Bajo+}	0,00373053
[30]	{reference=+Bajo+,cite=+Bajo+}	0,00368492

Tabla 7.6: Estudio 4 resultados de Eclat.

ID	X=>Y	Soporte	Confidencia	Elevación
[1]	{italic=-Alto-}=>{internal=-Alto-}	0,0029	0,82068276	94,05664
[2]	{heading2=-Alto-}=>{bulletedelement=-Alto-}	0,00158	0,80456758	132,92482
[3]	{external=-Alto-}=>{indent1=-Alto-}	0,00192	0,79679373	150,12876
[4]	{indent1=-Alto-}=>{internal=-Alto-}	0,00422	0,79543123	91,16262
[5]	{external=-Alto-}=>{internal=-Alto-}	0,0019	0,78828856	90,34402
[6]	{external=-Alto-}=>{bulletedelement=-Alto-}	0,00189	0,7870648	130,03314
[7]	{italic=+Alto+}=>{internal=+Alto+}	0,00277	0,76948238	93,73189
[8]	{external=+Alto+}=>{internal=+Alto+}	0,00179	0,75399282	91,84508
[9]	{indent1=+Alto+}=>{internal=+Alto+}	0,00394	0,75181271	91,57952
[10]	{external=+Alto+}=>{bulletedelement=+Alto+}	0,00179	0,75114523	105,27615
[11]	{external=+Alto+}=>{indent1=+Alto+}	0,00178	0,74904049	143,07084
[12]	{heading2=+Alto+}=>{bulletedelement=+Alto+}	0,00134	0,73584602	103,1319
[13]	{heading3=-Alto-}=>{bulletedelement=-Alto-}	0,00123	0,72665049	120,05193
[14]	{italic=-Alto-}=>{indent1=-Alto-}	0,00255	0,72102038	135,85185
[15]	{italic=-Alto-}=>{bulletedelement=-Alto-}	0,00248	0,70113793	115,83693
[16]	{italic=+Alto+}=>{indent1=+Alto+}	0,00247	0,68627852	131,08296
[17]	{italic=+Alto+}=>{bulletedelement=+Alto+}	0,00246	0,6833347	95,77222
[18]	{external=-Alto-}=>{italic=-Alto-}	0,00164	0,68292235	193,46355
[19]	{heading3=+Alto+}=>{bulletedelement=+Alto+}	0,00118	0,66039138	92,55662
[20]	{heading2=-Alto-}=>{internal=-Alto-}	0,00129	0,65511045	75,08077
[21]	{includeonly=-Alto-}=>{blockquote=-Alto-}	0,00143	0,65304068	185,46227
[22]	{external=+Alto+}=>{italic=+Alto+}	0,00155	0,65154141	181,04947
[23]	{heading2=-Alto-}=>{indent1=-Alto-}	0,00126	0,64275552	121,10549
[24]	{includeonly=+Alto+}=>{blockquote=+Alto+}	0,0017	0,6238656	109,34069
[25]	{reference=-Alto-}=>{internal=-Alto-}	0,00174	0,62032563	71,09415
[26]	{reference=-Alto-}=>{indent1=-Alto-}	0,00172	0,6134979	115,59288
[27]	{reference=-Alto-}=>{bulletedelement=-Alto-}	0,0017	0,60798319	100,44658
[28]	{heading2=-Alto-}=>{italic=-Alto-}	0,00117	0,59333583	168,08479
[29]	{blod=+Medio+}=>{internal=+Alto+}	0,00136	0,57196689	69,67221
[30]	{reference=-Alto-}=>{italic=-Alto-}	0,00159	0,56722689	160,68844

Tabla 7.7: Estudio 4 resultados de Apriori.

junto de los estilos como una regla real.

Además podemos observar que la gran mayoría de los cambios son de la medida "Alto". Esto se debe a que al disminuir el soporte se dejó espacio a las reglas relacionadas con los actos de vandalismo de eliminación masiva y los de recuperación de la información. En estas revisiones los estilos siempre se modifican aproximadamente de la misma forma por lo cual obtienen un alto grado de confianza y elevación.

7.6. Estudio 5 - Cluster Analysis y minería de reglas de asociación

7.6.1. Nombre

CAPA_PatronesEnClusters_7

7.6.2. Objetivo del estudio

En este estudio se busca a partir de los clusters hallados en *CA_RevisionesEvolEC_2* y del mecanismo de obtención de reglas utilizado en *PA_EstilosNivelados_6*, obtener información sobre qué reglas se aplican a cada uno de los clusters. A fin de que se puedan utilizar para brindar una mejor descripción o caracterización de los tipos de revisiones detectados. También con esto se espera poder obtener un conjunto de reglas con menos valores extremos que sesguen los resultados del minado de reglas de asociación obtenido en *PA_EstilosNivelados_6*.

7.6.3. ¿Con qué herramientas se realizaron?

A partir de la estructura de clusters obtenida en *CA_RevisionesEvolEC_2* se continuará trabajando con herramientas de clustering a fin de consolidar grupos a los que se les aplicará las herramientas de minado de reglas de asociación utilizadas en *PA_EstilosNivelados_6*.

7.6.4. Metodología y configuraciones

Se inició con el conjunto de clusters obtenidos en *CA_RevisionesEvolEC_2*. Por la problemática observada en *PA_EstilosNivelados_6* sobre las reglas que surgen de revisiones extremas derivadas de actos de vandalismo o su recuperación se decidió no solo

Cluster 1	Cluster 2	Cluster 3
0.9982468	0.6966620	0.4949600

Tabla 7.8: Estabilidad de 3 clusters segunda iteración.

utilizar este primer nivel de clustering sino que aislar aún más lo que se consideran revisiones normales podando los valores extremos de las mismas para posteriormente asignarlas a los clusters correspondientes de vandalismo o recuperación.

Para esto se tomó de los 3 clusters iniciales el cluster que agrupa la mayor cantidad de revisiones que fue identificado como el de ediciones normales y aplicar nuevamente el procedimiento de agrupación con K-means y 3 clusters. Con esto obtuvimos en uno de los 3 clusters un conjunto de revisiones muy inestable, según la métrica del coeficiente de Jaccard.

Por lo cual se procedió a sacarlo del conjunto de datos y a los restantes aplicarles nuevamente el procedimiento de agrupación con K-means y 3 clusters. Con esto obtuvimos en uno de los 3 clusters un conjunto de revisiones muy estable, según la métrica del coeficiente de Jaccard, que en este estudio representa a las revisiones normales dentro de los artículos.

Luego al conjunto de datos inicial se le extrajo el conjunto de revisiones normales, y a las revisiones restantes se les aplicó el algoritmo de agrupación K-means en esta ocasión con 2 clusters generando dos grupos bien definidos y muy estables según la métrica del coeficiente de Jaccard en los cuales puede observarse como uno representa las revisiones de eliminación extrema o actos de vandalismo mientras las otras representan las recuperaciones de los mismos.

Ya con estos 3 conjuntos bien definidos se procedió a aplicarles los algoritmos de reglas de asociación Eclat y Apriori con diversas configuraciones.

7.6.5. Resultados y análisis

En primera instancia presentaremos los resultados de los análisis de clustering realizados. Partiendo del conjunto inicial de 3 clusters que teníamos en *CA_RevisionesEvolEC_2* de los cuales en la Tabla 7.2 podemos observar la estabilidad que tenían. A partir de la segunda ejecución de K-means con 3 clusters utilizando el cluster de revisiones normales que tenia 0.99 nivel de estabilidad según la métrica del coeficiente de Jaccard se obtuvieron 3 clusters con la estabilidad que se muestra en la Tabla 7.8.

Cluster 1	Cluster 2	Cluster 3
0.7323271	0.9964999	0.4804540

Tabla 7.9: Estabilidad de 3 clusters tercera iteración.

Cluster 1	Cluster 2
0.9319298	0.9257944

Tabla 7.10: Estabilidad de 2 clusters cuarta iteración.

Del resultado anterior se tomaron solo los clusters 1 y 2 que tenían un nivel de estabilidad aceptable y se volvió a ejecutar el algoritmo con la misma configuración, el resultado de la estabilidad de los clusters obtenidos se muestra en la Tabla 7.9.

Al observar que la estabilidad del cluster de revisiones normales se era menor que en la iteración anterior se almaceno este grupo y se lo eliminó del conjunto inicial para aplicarle el algoritmo K-means con 2 clusters al conjunto restante. La estabilidad de los clusters resultantes se presenta en la Tabla 7.10. El cluster 1 representa las eliminaciones masivas mientras que el cluster 2 las aplicaciones masivas de estilos.

Ya con los 3 grupos bien definidos y con mejores niveles del coeficiente de Jaccard del que se había obtenido en el primer resultado del estudio *CA_RevisionesEvolEC_2* se continuó con el minado de reglas de los mismos.

Para el grupo de eliminaciones masivas primero se generaron sus respectivas transacciones y luego se aplicaron los algoritmos Eclat y Apriori. Eclat se ejecutó con un soporte mínimo de 0.1 y una cantidad mínima de elementos igual a 2.

Se obtuvo como resultado un total de 375 elementos de los cuales se presentan los 20 primeros ordenados por su soporte en la Tabla 7.11.

Como se puede observar en el resultado de Eclat y como era esperado se presentan múltiples reglas relacionadas con la eliminación en gran escala de los estilos. Además podemos observar el gran nivel de soporte llegando a un pico de 0,45.

Apriori se ejecutó con un soporte mínimo de 0.1, una confidencia mínima de 0.01 y una cantidad mínima de elementos de 2. Con esta configuración Apriori nos brindo 88 reglas de las cuales se presentan las primeras 20 ordenadas por su confidencia en la Tabla 7.12.

Como podemos observar en los resultados de Apriori para este grupo tenemos reglas

ID	Estilos	Soporte
[1]	{internal=-Alto-,indent1=-Alto-}	0,455696
[2]	{internal=-Alto-,bulletedelement=-Alto-}	0,392405
[3]	{bulletedelement=-Alto-,indent1=-Alto-}	0,367089
[4]	{italic=-Alto-,internal=-Alto-}	0,35443
[5]	{internal=-Alto-,bulletedelement=-Alto-,indent1=-Alto-}	0,341772
[6]	{italic=-Alto-,internal=-Alto-,indent1=-Alto-}	0,329114
[7]	{italic=-Alto-,indent1=-Alto-}	0,329114
[8]	{heading3=-Medio-,internal=-Alto-}	0,329114
[9]	{heading3=-Medio-,indent1=-Alto-}	0,303798
[10]	{italic=-Alto-,internal=-Alto-,bulletedelement=-Alto-}	0,291139
[11]	{italic=-Alto-,bulletedelement=-Alto-}	0,291139
[12]	{heading3=-Medio-,internal=-Alto-,indent1=-Alto-}	0,291139
[13]	{italic=-Alto-,heading3=-Medio-,internal=-Alto-}	0,278481
[14]	{italic=-Alto-,internal=-Alto-,bulletedelement=-Alto-,indent1=-Alto-}	0,278481
[15]	{italic=-Alto-,bulletedelement=-Alto-,indent1=-Alto-}	0,278481
[16]	{italic=-Alto-,heading3=-Medio-}	0,278481
[17]	{heading3=-Medio-,bulletedelement=-Alto-}	0,278481
[18]	{internal=-Alto-,external=-Alto-}	0,253165
[19]	{external=-Alto-,indent1=-Alto-}	0,253165
[20]	{heading2=-Medio-,internal=-Alto-}	0,253165

Tabla 7.11: Estudio 5 resultados de Eclat para el grupo de eliminación

ID	X=>Y	Soporte	Confidencia	Elevación
[1]	{blockquote=-Alto-}=>{includeonly=-Alto-}	0,151899	1	5,642857
[2]	{italic=-Alto-}=>{internal=-Alto-}	0,35443	1	1,795455
[3]	{external=-Alto-}=>{indent1=-Alto-}	0,253165	0,952381	1,835075
[4]	{external=-Alto-}=>{internal=-Alto-}	0,253165	0,952381	1,709957
[5]	{blod=-Medio-}=>{internal=-Alto-}	0,21519	0,944444	1,695707
[6]	{italic=-Alto-}=>{indent1=-Alto-}	0,329114	0,928571	1,789199
[7]	{heading4=-Medio-}=>{internal=-Alto-}	0,151899	0,923077	1,657343
[8]	{heading3=-Alto-}=>{bulletedelement=-Alto-}	0,113924	0,9	1,58
[9]	{big=-Alto-}=>{blockquote=-Alto-}	0,101266	0,888889	5,851852
[10]	{big=-Alto-}=>{includeonly=-Alto-}	0,101266	0,888889	5,015873
[11]	{infobox=-Bajo-}=>{indent1=-Alto-}	0,101266	0,888889	1,712737
[12]	{infobox=-Bajo-}=>{internal=-Alto-}	0,101266	0,888889	1,59596
[13]	{external=-Medio-}=>{internal=-Alto-}	0,202532	0,888889	1,59596
[14]	{blod=-Medio-}=>{indent1=-Alto-}	0,202532	0,888889	1,712737
[15]	{indent1=-Alto-}=>{internal=-Alto-}	0,455696	0,878049	1,576497
[16]	{heading2=-Alto-}=>{bulletedelement=-Alto-}	0,227848	0,857143	1,504762
[17]	{reference=-Alto-}=>{internal=-Alto-}	0,227848	0,857143	1,538961
[18]	{heading4=-Medio-}=>{italic=-Alto-}	0,139241	0,846154	2,387363
[19]	{heading4=-Medio-}=>{indent1=-Alto-}	0,139241	0,846154	1,630394
[20]	{heading3=-Medio-}=>{internal=-Alto-}	0,329114	0,83871	1,505865

Tabla 7.12: Estudio 5 resultados de Apriori para el grupo de eliminación.

también relacionadas con la eliminación de estilos y con un alto nivel de confidencia, pero sin embargo si observamos los niveles de elevación podemos ver que en su mayoría son relativamente bajos esto nos permite denotar el hecho de que no son reglas que reflejan principalmente patrones sino que su alto nivel de confidencia y ocurrencia derivan de ser parte de un conjunto de eliminaciones masivas siendo este claramente el principal descriptivo para la representación de este conjunto.

Continuando con el análisis de los diferentes clusters, el cluster que refleja aplicaciones en masa o recuperaciones del contenido fue convertido a transacciones para luego ser analizado con Eclat y Apriori.

Para este conjunto también Eclat fue ejecutado con un soporte mínimo de 0.1 y una cantidad mínima de elementos igual a 2.

Se obtuvo como resultado un total de 226 elementos de los cuales se presentan los 20 primeros ordenados por su soporte en la Tabla 7.13.

Como se puede observar en el resultado de Eclat los elementos con mayor ocurrencia tienen medidas de aplicación altas tal como era esperado para este grupo. Además se presentan soportes muy altos llegando a 0.46.

Junto con esto Apriori se ejecutó con un soporte mínimo de 0.1, una confidencia mínima de 0.01 y una cantidad mínima de elementos de 2. De esta ejecución Apriori nos retorno 64 reglas de las cuales se presentan las primeras 20 ordenadas por su confidencia en la Tabla 7.14.

Como podemos observar en el resultado de esta ejecución de Apriori ,y al igual que en la ejecución de Apriori con el grupo de eliminaciones masivas, aunque tenemos un alto nivel de confianza el la elevación nos indican que estas reglas no son patrones. Por lo cual podemos decir que en este grupo su mayor caracterización es el nivel de aplicaciones el cual es muy alto.

Por último un detalle importante a observar es que las reglas encontradas a diferencia de en otras ejecuciones de Apriori no suelen coincidir exactamente en el nivel de ocurrencia de X respecto a Y. Esto se debe principalmente a que la recuperación en masa de los elementos no conlleva necesariamente a que la aplicación de los mismos sea en grandes cantidades, por ejemplo, en la regla 18, en el caso de los infobox que son una estructura de los artículos de Wikipedia no suele haber más de uno por artículo por lo cual es esperable que al recuperarlo de su eliminación su nivel de co ocurrencia con otros estilos también sea alta aunque el nivel en que varíe su aplicación sea bajo.

Finalizando con los grupos se transformó en transacciones al grupo de las revisiones consideradas normales para luego aplicarles Eclat y Apriori. Eclat se ejecutó con 0.01

ID	Estilos	Soporte
[1]	{internal=+Alto+,indent1=+Alto+}	0,464789
[2]	{internal=+Alto+,bulletedelement=+Alto+}	0,450704
[3]	{internal=+Alto+,bulletedelement=+Alto+,indent1=+Alto+}	0,380282
[4]	{bulletedelement=+Alto+,indent1=+Alto+}	0,380282
[5]	{italic=+Alto+,internal=+Alto+}	0,352113
[6]	{italic=+Alto+,bulletedelement=+Alto+}	0,338028
[7]	{italic=+Alto+,internal=+Alto+,indent1=+Alto+}	0,309859
[8]	{italic=+Alto+,internal=+Alto+,bulletedelement=+Alto+}	0,309859
[9]	{italic=+Alto+,indent1=+Alto+}	0,309859
[10]	{italic=+Alto+,internal=+Alto+,bulletedelement=+Alto+,indent1=+Alto+}	0,295775
[11]	{italic=+Alto+,bulletedelement=+Alto+,indent1=+Alto+}	0,295775
[12]	{internal=+Alto+,external=+Alto+,indent1=+Alto+}	0,267606
[13]	{internal=+Alto+,external=+Alto+}	0,267606
[14]	{external=+Alto+,indent1=+Alto+}	0,267606
[15]	{heading2=+Medio+,internal=+Alto+,indent1=+Alto+}	0,253521
[16]	{heading2=+Medio+,bulletedelement=+Alto+}	0,253521
[17]	{heading2=+Medio+,internal=+Alto+}	0,253521
[18]	{heading2=+Medio+,indent1=+Alto+}	0,253521
[19]	{external=+Alto+,bulletedelement=+Alto+}	0,253521
[20]	{italic=+Alto+,external=+Alto+,bulletedelement=+Alto+}	0,239437

Tabla 7.13: Estudio 5 resultados de Eclat para el grupo de aplicación

ID	X=>Y	Soporte	Confidencia	Elevación
[1]	{external=+Medio+}=>{internal=+Alto+}	0,15493	1	1,775
[2]	{blod=+Alto+}=>{internal=+Alto+}	0,140845	1	1,775
[3]	{indent1=+Alto+}=>{internal=+Alto+}	0,464789	1	1,775
[4]	{external=+Medio+}=>{indent1=+Alto+}	0,140845	0,909091	1,955923
[5]	{external=+Medio+}=>{bulletedelement=+Alto+}	0,140845	0,909091	1,501057
[6]	{external=+Alto+}=>{indent1=+Alto+}	0,267606	0,904762	1,946609
[7]	{external=+Alto+}=>{internal=+Alto+}	0,267606	0,904762	1,605952
[8]	{cite=+Alto+}=>{bulletedelement=+Alto+}	0,126761	0,9	1,486047
[9]	{blod=+Alto+}=>{indent1=+Alto+}	0,126761	0,9	1,936364
[10]	{blod=+Alto+}=>{bulletedelement=+Alto+}	0,126761	0,9	1,486047
[11]	{heading2=+Medio+}=>{indent1=+Alto+}	0,253521	0,9	1,936364
[12]	{heading2=+Medio+}=>{internal=+Alto+}	0,253521	0,9	1,5975
[13]	{heading2=+Medio+}=>{bulletedelement=+Alto+}	0,253521	0,9	1,486047
[14]	{italic=+Alto+}=>{internal=+Alto+}	0,352113	0,862069	1,530172
[15]	{blod=+Bajo+}=>{internal=+Alto+}	0,169014	0,857143	1,521429
[16]	{external=+Alto+}=>{bulletedelement=+Alto+}	0,253521	0,857143	1,415282
[17]	{numberedelement=+Bajo+}=>{internal=+Alto+}	0,140845	0,833333	1,479167
[18]	{infobox=+Bajo+}=>{italic=+Alto+}	0,140845	0,833333	2,04023
[19]	{italic=+Alto+}=>{bulletedelement=+Alto+}	0,338028	0,827586	1,36648
[20]	{external=+Medio+}=>{heading2=+Medio+}	0,126761	0,818182	2,904546

Tabla 7.14: Estudio 5 resultados de Apriori para el grupo de aplicación

ID	Estilos	Soporte
[1]	{blockquote=+Bajo+,includeonly=+Bajo+}	0,021421
[2]	{internal=+Bajo+,indent1=+Bajo+}	0,016751
[3]	{internal=+Bajo+,bulletedelement=+Bajo+}	0,013096
[4]	{heading2=+Bajo+,includeonly=+Bajo+}	0,008122
[5]	{s=+Bajo+,includeonly=+Bajo+}	0,007411
[6]	{external=+Bajo+,indent1=+Bajo+}	0,00731
[7]	{reference=+Bajo+,indent1=+Bajo+}	0,00731
[8]	{bulletedelement=+Bajo+,indent1=+Bajo+}	0,007208
[9]	{blockquote=+Bajo+,big=+Bajo+}	0,006904
[10]	{reference=+Bajo+,bulletedelement=+Bajo+}	0,006497
[11]	{heading3=+Bajo+,bulletedelement=+Bajo+}	0,005685
[12]	{external=+Bajo+,bulletedelement=+Bajo+}	0,005584
[13]	{heading2=+Bajo+,indent2=+Bajo+}	0,005381
[14]	{blockquote=+Bajo+,cite=+Bajo+}	0,005076
[15]	{heading2=+Bajo+,bulletedelement=+Bajo+}	0,005076
[16]	{internal=-Medio-,bulletedelement=-Bajo-}	0,004873
[17]	{big=+Bajo+,includeonly=+Bajo+}	0,00467
[18]	{blockquote=+Medio+,includeonly=+Bajo+}	0,004365
[19]	{sub=+Bajo+,big=+Bajo+}	0,004365
[20]	{blockquote=+Bajo+,bulletedelement=+Bajo+}	0,004365

Tabla 7.15: Estudio 5 resultados de Eclat para el grupo estándar

como soporte mínimo y una cantidad mínima de 2 elementos por retorno. Con esta configuración se obtuvieron 251 resultados de los cuales se presentan los primero 20 ordenados según su soporte en la Tabla 7.15.

Como podemos observar en los resultados de Eclat los elementos de mayor soporte ya no están ocultos u opacados por los eventos aislados de eliminaciones en masa o recuperación de las eliminaciones. Además podemos observar que aunque se presentan elementos que denotan disminución en los estilos aplicados, hay una tendencia al crecimiento y el aumento de los estilos aplicados. Esto refleja la tendencia común de los artículos a crecer además de denotar que en ese proceso es donde ocurre la mayor integración de reglas y patrones respecto a cómo estructurar el artículo con la aplicación

ID	X=>Y	Soporte	Confidencia	Elevación
[1]	{indent2=+Bajo+}=>{heading2=+Bajo+}	0,005381	0,588889	25,78025
[2]	{s=+Bajo+}=>{includeonly=+Bajo+}	0,007411	0,58871	13,06034
[3]	{blockquote=+Alto+}=>{includeonly=+Medio+}	0,002132	0,567568	68,17732
[4]	{s=-Medio-}=>{includeonly=-Medio-}	0,001421	0,518519	62,28546
[5]	{blod=-Medio-}=>{internal=-Alto-}	0,001117	0,5	104,7872
[6]	{blockquote=+Alto+}=>{italicblod=+Bajo+}	0,001827	0,486486	13,7698
[7]	{includeonly=+Bajo+}=>{blockquote=+Bajo+}	0,021421	0,475225	8,343972
[8]	{reference=-Medio-}=>{indent1=-Medio-}	0,002843	0,451613	39,71774
[9]	{italic=-Medio-}=>{indent1=-Medio-}	0,002538	0,446429	39,2618
[10]	{external=-Medio-}=>{indent1=-Medio-}	0,001218	0,444444	39,0873
[11]	{heading2=+Medio+}=>{includeonly=+Medio+}	0,001421	0,4375	52,55335
[12]	{internal=-Alto-}=>{indent1=-Medio-}	0,00203	0,425532	37,42401
[13]	{cite=-Medio-}=>{reference=-Medio-}	0,001421	0,424242	67,3998
[14]	{heading2=-Medio-}=>{includeonly=-Medio-}	0,001523	0,416667	50,05081
[15]	{indent2=+Bajo+}=>{includeonly=+Bajo+}	0,003655	0,4	8,873874
[16]	{internal=+Alto+}=>{indent1=+Medio+}	0,001726	0,395349	56,43748
[17]	{italic=+Medio+}=>{internal=+Alto+}	0,001523	0,394737	90,42228
[18]	{s=-Bajo-}=>{includeonly=-Bajo-}	0,002335	0,383333	30,69783
[19]	{external=+Bajo+}=>{indent1=+Bajo+}	0,00731	0,380952	8,470386
[20]	{indent1=+Bajo+}=>{internal=+Bajo+}	0,016751	0,372461	6,744

Tabla 7.16: Estudio 5 resultados de Apriori para el grupo estándar

de diversos estilos en conjunto.

También se ejecutó Apriori con una configuración de 0.01 de soporte y confidencia mínimos, y con 2 como mínimo de elementos por regla obtenida. Con esto se obtuvieron un total de 215 elementos de los cuales se presentan los primeros 20 elementos ordenados según su confidencia en la Tabla 7.16.

Con los resultados de Apriori podemos ver que que a diferencia de los dos grupos anteriores en este caso las reglas no solo tienen una confidencia alta llegando a un nivel de 0.58 sino que también todas presenta una elevación mucho mayor a la de una situación aleatoria, cuyo valor sería uno. Por lo cual podemos decir que en este caso estamos

probablemente frente a casos en los que las reglas halladas sean realmente patrones.

De este estudio por lo tanto obtuvimos caracterizaciones para los dos tipos de los 3 tipos de clusters hallados en el estudio *CA_RevisionesEvolEC_2* y también obtuvimos un conjunto de reglas más limpio y aplicable a lo que se consideran revisiones normales realizadas en el día a día por los editores de Wikipedia en los artículos.

Capítulo 8

Conclusiones

8.1. Aportes realizados

En este trabajo se presenta un marco teórico sobre el estado del arte del estudio de wikis en la actualidad y sobre estudios de visualización de datos, proveniencia de la información y minería de patrones. Se analizan diversos trabajos realizados sobre MediaWikis brindando una comparación con los trabajos propuestos y cómo estos complementan, expanden o se fundamentan en los mismos.

Se presenta un análisis referente a las fuentes de información y como se encuentra estructurada la misma. Se analizan dos fuentes de información en profundidad, los dumps de MediaWikis de los cuales se explica la forma de obtención y los diversos tipos de información contenida en los mismos. También se estudia la obtención dinámica de artículos de MediaWiki con sus respectivas revisiones. Y por último se plantean las ventajas y desventajas de cada fuente de información.

También se presentan un conjunto de métricas para evaluar la información obtenida y un enfoque basado en el lenguaje de marcado utilizado por las wikis para la organización, estructuración y formato de su contenido. De estas métricas y este enfoque particular se desarrolla un prototipo para la obtención de información de las fuentes mencionadas el cual se aplica particularmente a la Wikipedia en inglés.

Para este desarrollo se brinda información sobre las diversas tecnologías utilizadas y el porqué de la selección de cada una.

El Prototipo permite descargar y obtener la información sobre los artículos seleccionados, esto incluye toda la información de sus revisiones y el procesamiento de las mismas para obtener los valores utilizados en las métricas ya desarrolladas. Además de la obtención y procesamiento de la información se generan gráficos referentes a la información obtenida los cuales permiten un análisis manual de los resultados obteni-

dos. La información para la obtención y instalación de este prototipo se encuentra en el Apéndice A.

A partir del prototipo también se presenta el análisis de un conjunto de datos acotados para posibilitar un análisis manual. Se presentan en primera instancia los resultados obtenidos directamente del prototipo y luego un análisis de algunos de los casos de interés más relevantes detectados. Para finalizar se explica la importancia de estos resultados y su relación con los sucesos reales que los generan además de los mismos se deriva un enfoque distinto para la detección de vandalismo en las revisiones a artículos de Wikipedia.

Sobre este prototipo y los resultados estudiados se realizó una ponencia en el congreso de humanidades digitales del año 2016. (31)

Para poder continuar los análisis de la información obtenida el prototipo provee también la posibilidad de exportar los datos obtenidos en formato json.

También se presentan R como lenguaje y plataforma especializada en análisis estadísticos y se presentan las técnicas de aprendizaje no supervisado para el desarrollo de los estudios. Las técnicas presentadas son kmeans como técnica de clustering y dos algoritmos para minado de patrones y reglas de asociación conocidos como Eclat y Apriori.

Luego se presentan el conjunto de datos a utilizar en los distintos análisis realizados con dichas técnicas, el cual consiste en la información de artículos y revisiones de la Wikipedia en inglés obtenidos con el primer prototipo pero a una escala más grande que las evaluadas con el mismo.

De los 5 estudios realizados se presenta:

- Nombre otorgado para identificar el caso de estudio.
- Información sobre que se desea estudiar o analizar.
- Los datos utilizados para el estudio, junto con qué métricas los generó o de qué información derivan.
- Desarrollo del caso de estudio con información sobre metodología,algoritmos y configuraciones utilizadas.
- Presentación y análisis de los resultados obtenidos.

De estos estudios podemos observar claramente reflejado un proceso de análisis en el que la información obtenida de un estudio es aplicada a los siguientes. Por esto este conjunto de estudios muestran indicios de cómo trabajar con la ciencia de los datos, y

junto con esto los resultados de los estudios entre los que se incluyen:

- Obtención de un conocimiento de cómo es la distribución y evolución de los datos analizados.
- Tratamiento de los datos para que los mismos revelen información de interés.
- Obtención de conocimiento sobre las limitaciones de cada una de las técnicas utilizadas y sobre cómo combinarlas para solventar dichas limitaciones.
- Sobre la información se obtiene una clasificación sobre los tipos de revisiones y una caracterización de las mismas.
- También se obtienen nociones sobre patrones reales del comportamiento de los editores durante las ediciones que llevan a cabo. Sobretodo en lo referente a las combinaciones estructurales o de estilo utilizadas como normativa implícita para organizar los artículos.
- Junto con los patrones sobre la relación de cambio en la estructura del artículo se obtiene información sobre en qué medida son desarrollados estos cambios. Dichas medidas nos permiten ver la relación entre los distintos cambios estructurales.

A continuación se presentan los trabajos que deberían realizarse para continuar con el análisis de los patrones en la evolución de las wikis y su aplicación al servicio de la comunidad.

8.2. Trabajos futuros

En este trabajo se utilizo un prototipo el cual obtenía parte de la información disponible dentro de las revisiones de las wikis, una posible tarea a realizar es extender dicho prototipo para que obtenga diferentes tipos de información como puede ser la extracción de contenido semántico de los artículos para ser analizados posteriormente. También el prototipo era orientado principalmente a artículos de wikis, y resulta de interés poder relacionar los cambios que suceden en los artículos con las paginas de charlas de los artículos en cuestión para poder obtener una mejor visión de la motivación de los cambios. También durante la etapa de análisis los estudios en los que se aplicaron algoritmos de clustering se realizaron utilizando la distancia Euclidiana por lo cual seria importante en un futuro aplicar los algoritmos con diferentes tipos de distancias como podrían ser la distancia de Hamming, la distancia Manhattan o utilizando similitud coseno. Para obtener una ventaja que permita realizar aportes significativos a las MediaWikis a partir de los patrones detectados deben generarse algoritmos que permitan la categorizacion automática de las diferentes tipos de revisiones detectadas en este trabajo. Para este fin se puede desarrollar un algoritmo que implementa técnicas

de aprendizaje supervisado que permite la clasificación de las revisiones a medida que son generadas siguiendo con las características representativas de las mismas obtenidas en este trabajo.

Además a partir de los estudios existentes en conjunto con este generar una integración del estudio de los artículos, las revisiones, los editores y su actividad a fin de obtener un ecosistema en el que se tengan en cuenta todas las variables que influyen en la creación de conocimiento. Y de estas variables poder predecir la información relevante para el tipo de actividad desarrollada por un usuario en un momento determinado a fin de suministrarles facilitando y brindando soporte a la tarea que se encuentre realizando.

Para ello además se requerirá un conjunto de fuentes de información confiables ya sean internas o externas a la comunidad generadora del conocimiento.

También como se planteó durante el trabajo se espera que puedan generarse nuevas herramientas para la detección de vandalismo en artículos de Wikipedia que posean entre sus variables de interés los cambios estructurales para mejorar la detección de casos de vandalismo en los que los sistemas actuales no logran detectarlos.

Apéndice A

Guías y contacto

A.1. Instalación del prototipo

A.1.1. Requerimientos

Para instalar el prototipo se requieren los siguientes software:

- Tomcat 7 server
- H2 Database Engine.
- Eclipse (Opcional).

A.1.2. Guía de instalación

1. El prototipo se encuentra en el repositorio WebWiki[https://github.com/jonx18/WebWiki].
2. Para ejecutar sobre el Tomcat server descargar el archivo WikiWebTest.war [https://github.com/
y realizar el despliegue de forma convencional.
3. Acceder al archivo "historyPath.properties" ubicado en /Tomcat/webapps/WikiWebTest/WEB-
INF/clases/.
4. Modificar el valor de la clave "export.path" con el de un directorio ya existente para que se exporten los resultados.
5. Agregar claves con valores que indiquen la localización de los archivos a procesar como se muestra en el mismo archivo.
6. Acceder a la interfaz web desde WikiWebTest [http://localhost:8080/WikiWebTest/] para comprobar el correcto funcionamiento del prototipo.

A.2. Scripts de estudios

Los código utilizados para realizar los análisis en la segunda parte de esta tesina requieren R version 3.3.2 para ejecutarse. Los mismos se encuentran en el repositorio RWiki [https://github.com/jonx18/RWiki].

A.3. Contacto

Jonathan Martin.

Email: jonamar10@hotmail.com ; jonamar10@gmail.com

Bibliografía

[1] Jose Felipe Ortega Soto. **Wikipedia: A quantitative analysis**, 2009. 1

[2] Emilio José Rodríguez Posada Noelia Sales Montes Manuel Palomo Duarte, Inmaculada Medina Bulo. **Tecnologías Wiki y conocimiento abierto en la universidad**, 2009. 1, 15

[3] M Valcke H Van Keer B De Wever, T Schellens. **Content analysis schemes to analyze transcripts of online asynchronous discussion groups**, 2005. 1

[4] Apurva A Desai Jatinderkumar R Saini. **A textual analysis of digits used for designing Yahoo-group identifiers**, 2010. 1

[5] Gerry Stahl. **Group Cognition: Computer Support for Building Collaborative Knowledge**, 2006. 2

[6] Kushal Dave Fernanda B. Viégas, Martin Wattenberg. **Studying Cooperation and Conflict between Authorswith history flow Visualizations**, 2004. 2, 15

[7] Mikka Ryokas. **A quote on the NY Times**. 5

[8] John Fletcher Ward Cunningham, Jeff Grigg. **Wiki History**. 6

[9] Larry Sanger. **The Early History of Nupedia and Wikipedia: A Memoir**, 2005/04/18. 6

[10] Friedman Vitaly. **Data Visualization and Infographics**, 2008/01/14. 12

[11] Mike Bostock. **D3 Data-Driven Documents**. 12

[12] Chart.js. **Chart.js**. 12

[13] Google. **Google Charts**. 12

[14] Olaf Hartig. **Provenance Information in the Web of Data**, 2009/04/20. 12

[15] John Mount Nina Zumel. **Practical Data Science with R**, 2014. 13, 74, 76

[16] D.S. Rajpoot Dr. Kanak Saxena. **A Way to Understand Various Patterns of Data Mining Techniques for Selected Domains** , 2009/05. 13

[17] Robe Kraut Eduard Hovy Diyi Yang, Aaron Halfaker. **Who Did What: Editor Role Identification in Wikipedia**, 2016. 14

[18] Alexandre Passant Fabrizio Orlandi. **Modelling Provenance of DBpedia Resources Using Wikipedia Contributions**, 2011. 15

[19] Aaron Halfaker R. Stuart Geiger. **Using Edit Sessions to Measure Participation in Wikipedia**, 2013. 15

[20] Ángela Ríos Luna. **Implementación del Patrón MVC en Aplicaciones Web con Java mediante la Integración de los Framework Hibernate, Spring y Primefaces**, 2014/07/02. 27

[21] Keisuke Nakano Alain Frisch. **Streaming XML transformations using term rewriting**, 2007/01/20. 27

[22] Robert Gentleman Ross Ihaka. **R Project**. 73

[23] Leo Osvald Jan Vitek Floreal Morandat, Brandon Hill. **Evaluating the Design of the R Language: Objects and Functions For Data Analysis**, 2012/06/11. 73

[24] Dan Toomey. **R for Data Science**, 2014. 73

[25] Peter J. Rousseeuw Leonard Kaufman. **Finding Groups in Data: An Introduction to Cluster Analysis**, 2008/05/27. 74

[26] Antonio Quintero Toscano Rafael Ayala Gomez, Eladio Dominguez. **Elementos de la Topologia General**, 2000/04/17. 75

[27] Javier Muguerza Jesús M. Pérez Iñigo Perona Olatz Arbelaitz, Ibai Gurrutxaga. **An extensive comparative study of cluster validity indices**, 2013/01. 75

[28] Peter J. Rousseeuw. **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis**, 1986. 75

[29] Lars Schmidt-Thieme. **Algorithmic Features of Eclat**, 2004/01. 76

[30] Ramakrishnan Srikant Rakesh Agrawal. **Fast Algorithms for Mining Association Rules**, 1994/09/12. 76

[31] Torres Diego Martin Jonathan. **Presentacion Análisis de Patrones en la Evolución de Wikis**. 104