Universitat Oberta de Catalunya

uoc.edu

Herramienta de privacidad para la realización de consultas web en Google

Judit Caballero Moro
Grado de Ingeniería Informática
Ingeniería de Computadores

Jorge Miguel Moneo
Helena Rifà Pous

03/06/2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	Herramienta de privacidad para la realización de consultas web en Google
Nombre del autor:	Judit Caballero Moro
Nombre del consultor/a:	Jorge Miguel Moneo
Nombre del PRA:	Helena Rifà Pous
Fecha de entrega:	06/2021
Titulación:	Grado de Ingeniería Informática
Área del Trabajo Final:	Seguridad Informática
Idioma del trabajo:	Español
Palabras clave	Consultas Web, Generalización, Privacidad, PET

Resumen del Trabajo

Este último año la tecnología ha tomado más relevancia que nunca, el número de usuarios crece de manera exponencial lo que dificulta mantener una seguridad adecuada para nuestros datos más sensibles. Si bien existen tecnologías que cumplen esta finalidad, no se dispone de herramientas capaces de evitar o distorsionar los perfiles de intereses creados a raíz de las búsquedas en internet, siendo uno de los servicios web más utilizados.

En este proyecto se marca como meta el desarrollo de una herramienta capaz de perturbar los datos recolectados por los motores de búsqueda, donde gracias a una investigación de mercado sobre qué mecanismos y tecnologías se encuentran disponibles, se toma la decisión de crear una extensión para Google Chrome que implemente el mecanismo de generalización de consultas.

Para conseguir nuestro objetivo, es esencial una tecnología que permita la relación semántica entre palabras, siendo WordNet nuestra base de conocimiento. Por otro lado, para comprobar que nuestra herramienta cumplía con las expectativas deseadas, se realizan una serie de pruebas con usuarios reales que permiten demostrar que la generalización escogida proporciona unos resultados acordes con la búsqueda original.

Por último, en base a las opiniones aportadas por los participantes y al análisis realizado a lo largo del proyecto, se llega a la conclusión que la herramienta diseñada supone un primer paso hacia unas consultas webs más seguras, sin comprometer los datos más sensibles de los usuarios a la vez que proporcionan búsquedas de interés.

Abstract:

Last year was one of the most relevant years for technology, the number of users is constantly growing which makes it difficult to have a proper security of our most sensitive data. Although there are technologies that help in this purpose, we cannot find any tool to avoid or distort the user profiles created from Internet searches, being one of the most popular web services.

The aim of this project is to develop a tool which disturbs the data collected by search engines. Thanks to market research, we found out what mechanisms and technologies are currently available, so it was decided to create an extension for Google Chrome that implements the query generalization mechanism.

To achieve our goal, a technology is necessary to allow the semantic relationship between words, choosing WordNet as our knowledge base. On the other hand, to verify the tool works exactly as expected, a series of tests are carried out with real users that allow us to demonstrate that the chosen generalization provides results in accordance with the original search.

Finally, based on the opinions contributed by the participants and the analysis done throughout the project, it is concluded that the designed tool is a first step towards secure web query, without compromising the most sensitive data of the users at the same time as providing searches of interest.

Índice

Índice de ilustraciones

Índice de tablas

1. Introducción

1.1. Problema por resolver

En los últimos años ha habido un crecimiento de las redes de manera exponencial, proporcionando una gran cantidad de servicios en la web como es el caso de las búsquedas por internet, las actividades comerciales o la creación de las redes sociales.

No obstante, este último año marcado por la pandemia ha conseguido que la red sea más relevante que nunca. El número de usuarios en internet en el mundo se ha incrementado un 7,3% respecto al año anterior, obteniendo los dispositivos móviles el mayor tráfico de red con un 4,3% más que en 2019 [1].

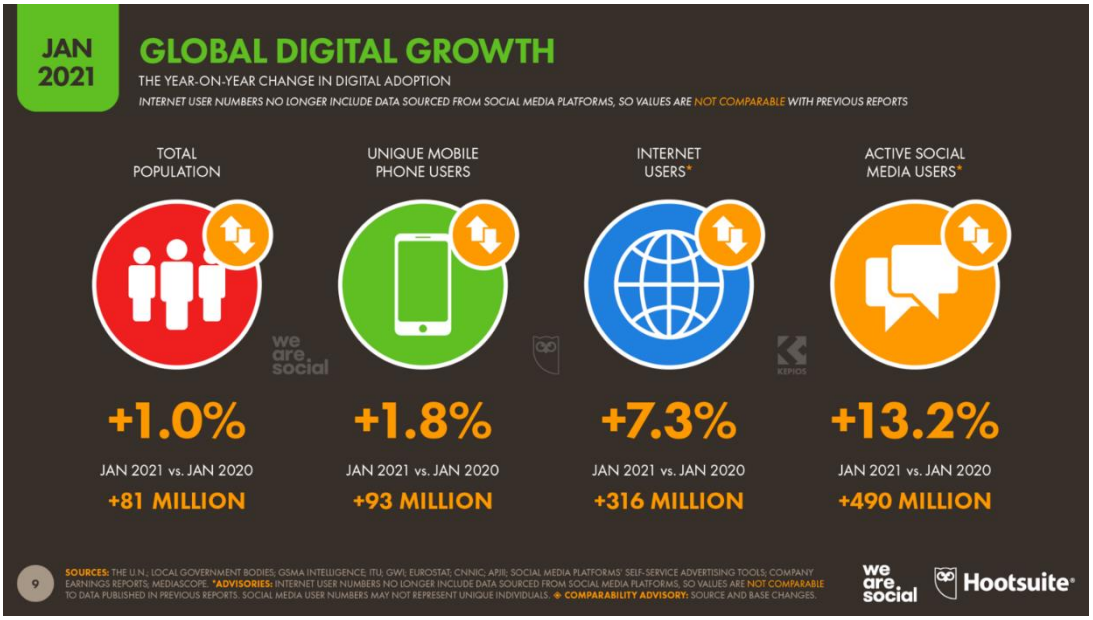


Ilustración 1 Fuente: S. Kemp (2021). Annual Global Digital Growth January 2021. Data Reportal. URL: https://datareportal.com/reports/digital-2021-global-overview-report

Este crecimiento supone una problemática en el ámbito de la seguridad informática, ya que al tener un mayor número de usuarios en la red hace indispensable aumentar el uso de mecanismos que otorguen al usuario una protección segura frente ataques maliciosos o suplantaciones de identidad y que hoy en día no disponemos.

Pero no solo ha crecido el número de usuarios, el teletrabajo, las clases online y las compras por internet ha supuesto un aumento en la realización de consultas que se llevan a cabo en motores de búsquedas como Google, Bing, Yahoo! o DuckDuckGo, posicionando a Google como el buscador más utilizado con un total de 92 millones de visitas en 2020.

Estos buscadores ofrecen consultas personalizadas para mejorar la experiencia del usuario, pero ¿Qué datos son los que almacenan y cuáles no? Su falta de transparencia provoca una inseguridad en el usuario causando una mayor necesidad de proteger su identidad y, por lo tanto, la utilización de un mecanismo a nivel de usuario que lo permita.

1.1.1. Relevancia del problema

Como se ha mencionado, los motores de búsqueda utilizan las consultas realizadas por los usuarios de Internet para la creación de perfiles personales, los cuales contienen gran cantidad de información confidencial presentando una amenaza para la privacidad. En el momento en que un usuario ha realizado una búsqueda, los motores almacenan en forma de registros todas las consultas enviadas dentro de sus bases de datos, gracias al uso de las cookies o a la configuración del propio navegador.

Por otro lado, el desconocimiento de dichos motores respecto a cómo utilizan nuestra información, hace necesario encontrar un mecanismo que permita seguir realizando búsquedas personalizadas, pero sin poner en riesgo la privacidad de uno mismo.

1.2. Estado de arte

En la actualidad existen varios motores de búsqueda para poder realizar consultas, entre los más conocidos encontramos Google, Yahoo!, Bing y DuckDuckGo. Tal como se ha comentado anteriormente, este último año ha sido clave para cada uno de ellos, puesto que su uso se ha visto aumentado. Estos buscadores ofrecen una mejor experiencia al usuario utilizando una serie de tecnologías para hacer sus consultas más rápidas y cercanas al usuario final, pero ¿Cómo realizan esta personalización de consultas?

La respuesta a dicha pregunta es mediante perfiles de usuarios. Estos perfiles se crean a partir de diferentes técnicas, una de las más conocidas consiste en almacenar información del usuario mediante el historial de navegación, ya sean por enlaces visitados o consultas anteriores o bien, haciendo uso de la ubicación del dispositivo. Gracias a todos estos datos, el buscador crea un perfil y muestra los resultados según sus preferencias de manera rápida y sencilla. También se puede hacer uso del perfil del usuario en redes sociales, por ejemplo, mirando el apartado de intereses que incluye Facebook, permite clasificar consultas según sus gustos. Pero ¿Qué sucede con esa información recolectada? ¿Están protegidos nuestros datos?

La falta de respuestas es lo que ha motivado a pensar nuevas formas que permitan ocultar aquella información más sensible. A pesar de que podemos encontrar algunas alternativas que nos proporcionen cierta privacidad, estas requieren la colaboración de agentes externos provocando una inseguridad al depender de su disponibilidad. Por ello, en este trabajo nos vamos a centrar en aquellos esquemas que funcionan directamente en la parte del usuario, concretamente aquellos que dan solución al problema generando y enviando consultas distorsionadas al motor de búsqueda sobre los intereses del usuario.

Actualmente existen varios esquemas en la literatura sobre métodos de generalización y falsificación de consultas, una propuesta la encontramos en [2], donde se crea un perfil local a partir de la red social Twitter. Este perfil se construye a partir de la recopilación de sus intereses con el fin de ser analizados semánticamente y clasificados en categorías, donde cada una de ellas recibirá una puntuación según si mantienen un interés mayor o menor en el usuario. Por otra parte, se crea un perfil público creado a partir de consultas anteriormente solicitadas

en el buscador con la intención de comparar ambos perfiles y detectar las diferencias para posteriormente crear la nueva consulta falsa.

Una segunda propuesta la podemos encontrar en [3] donde los autores presentan la generalización de consultas sin estar vinculado a un protocolo particular. En este caso, el grado de distorsión puede variar según la distancia semántica o indicando la cantidad de consultas falsas que se quieran enviar. Como sucedía en el caso anterior, se analizan consultas ya enviadas para generar nuevas relacionadas semánticamente, pero en este caso la interpretación semántica de datos textuales se basa en evidencias encontradas en una o varias fuentes de conocimiento construido, concretamente WordNet y ODP. Las consultas de usuarios se asignan a conceptos haciendo coincidir sus etiquetas textuales, una vez recibe la consulta del usuario aplica el análisis morfosintáctico y las clasifica según la información proporcionada. Por último, se construye la nueva consulta que se relaciona semánticamente con los principales temas mencionados por la original.

En este proyecto se seguirá un esquema parecido al último planteado, donde se obliga al motor de búsqueda a crear un perfil de usuario más orientado a los intereses generales del usuario y no tanto a detalles más sensible, mezclando consultas reales con falsas creadas a partir de la generalización de estas.

1.3. Objetivos del Trabajo

Para dar una solución a la problemática anterior, este proyecto tendrá como objetivo la creación y el análisis de una herramienta de privacidad que permita a los usuarios de internet proteger sus datos más confidenciales en el momento de realizar una consulta.

Objetivos generales:

- Diseño de una herramienta de privacidad cuya función sea proteger los datos privados de los usuarios en un motor de búsqueda.
- Testeo de la herramienta con usuarios reales para evaluar su usabilidad y rendimiento.

Objetivos específicos:

- Estudio de los diferentes mecanismos existentes para la protección de datos privados.
- Investigación de tecnologías de mejora de la privacidad disponibles a nivel de usuario.
- Selección del mecanismo de protección a implementar.
- Uso de WordNet como base de conocimiento para la herramienta.

1.4. Enfoque y método seguido

La metodología aplicada tendrá como fin el cumplimiento de los objetivos, se dividirá en dos partes: una parte de investigación y una parte práctica llevada a cabo mediante una metodología ágil, donde se llevará a cabo el diseño y desarrollo de una nueva herramienta.

Primero de todo, el estudio estará centrado en conocer los diferentes mecanismos de protección de privacidad que existen en la actualidad. Así como, el estudio de

mercado respecto a tecnologías que puedan proporcionar una protección de datos a nivel de usuario. Es importante conocer qué mecanismos se encuentran en uso y cuales proporcionan mejor resultado, puesto que uno de ellos será utilizado en el diseño de la tecnología.

Una vez realizada la investigación y el análisis de mercado sobre tecnologías de privacidad, se habrá adquirido la información necesaria para la implementación del mecanismo y los requisitos principales que deberá cumplir la herramienta dando paso a su desarrollo.

La segunda parte se realizará mediante un proceso dividido en fases. Estas fases serán el diseño, desarrollo y evaluación de la herramienta y se llevarán a cabo de forma iterativa. En este caso, se utilizará WordNet como base de conocimiento, ya que consiste en una tecnología que permite la generalización de consultas. Su creación estará enfocada para ser útil en el motor de búsqueda Google, puesto que es el que cuenta con el mayor número de usuarios.

Una vez el mecanismo sea completamente funcional, se realizará un testeo con un conjunto de participantes reales para analizar la percepción que tienen dichos usuarios respecto a sus consultas junto con el rendimiento de la herramienta en términos de consumo de CPU y memoria, así como la velocidad de respuesta.

1.5. Planificación del Trabajo

La metodología expuesta se cumplirá mediante una serie de tareas que corresponden a cada uno de los pasos a seguir para el buen desarrollo del trabajo.

A continuación, se detallan cada una de las tareas que forman parte del proyecto:

a) Investigación de mecanismos de protección de privacidad.

Esta tarea tiene como objetivo conocer los diferentes mecanismos de protección de privacidad que existen en la actualidad con el fin de demostrar por qué la generalización de consultas es un buen mecanismo para nuestro objetivo.

b) Estudio de tecnologías para proteger los datos.

En esta tarea se pretende encontrar que tipo de tecnologías existen para proteger la identidad de un usuario en la red y así demostrar la carencia de herramientas en el ámbito de consultas web.

c) Estudio de la herramienta WordNet.

Anteriormente se ha mencionado WordNet como la base de conocimiento del proyecto, por ello es conveniente realizar un estudio previo que explique en qué consiste y cómo está formada e implementada esta tecnología.

d) Desarrollo de una herramienta de privacidad.

Una vez dispongamos de toda la información necesaria para la creación de una nueva herramienta, se dará paso a su diseño e implementación.

e) Pruebas de la herramienta con usuarios reales y evaluación del rendimiento.

La tarea de realización de pruebas tiene como fin descubrir la viabilidad del propio mecanismo.

f) Redacción de la memoria y conclusiones.

Una vez realizadas todas las tareas anteriores únicamente quedará acabar de redactar la memoria del proyecto, donde recoja cada uno de los procesos

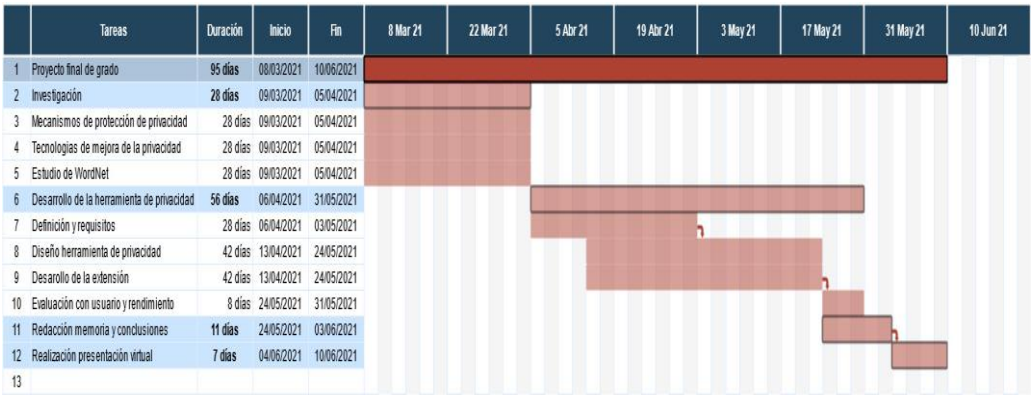
previamente explicados junto con las conclusiones obtenidas a lo largo del proceso.

g) Realización de la presentación virtual.

La última tarea consiste en preparar una presentación visual del proyecto donde se dé a conocer la metodología utilizada, la funcionalidad de nuestra herramienta y el aprendizaje obtenido.

1.6. Planificación temporal

La planificación temporal creada tiene como referencia las cuatro fechas claves para las entregas de las diferentes PECs.



	Tareas	Duración	Inicio	Fin
1	Proyecto final de grado	95 días	08/03/2021	10/06/2021
2	Investigación	28 días	09/03/2021	05/04/2021
3	Mecanismos de protección de privacidad	28 días	09/03/2021	05/04/2021
4	Tecnologías de mejora de la privacidad	28 días	09/03/2021	05/04/2021
5	Estudio de WordNet	28 días	09/03/2021	05/04/2021
6	Desarrollo de la herramienta de privacidad	56 días	06/04/2021	31/05/2021
7	Definición y requisitos	28 días	06/04/2021	03/05/2021
8	Diseño herramienta de privacidad	42 días	13/04/2021	24/05/2021
9	Desarollo de la extensión	42 días	13/04/2021	24/05/2021
10	Evaluación con usuario y rendimiento	8 días	24/05/2021	31/05/2021
11	Redacción memoria y conclusiones	11 días	24/05/2021	03/06/2021
12	Realización presentación virtual	7 días	04/06/2021	10/06/2021
13				

Ilustración 2 Diagrama de Gantt del proyecto

1.6.1. Fechas significativas y observaciones

En la siguiente tabla se muestran las fechas significativas resultantes de la planificación anterior. También se incluye una explicación final de los posibles contratiempos que puede haber.

Fecha	**Descripción**
08/03/2021	Definición del proyecto a llevar a cabo junto con el primer borrador del plan de trabajo.
05/04/2021	Primera entrega correspondiente a toda la parte de investigación que incluye el proyecto.
03/05/2021	Segunda entrega donde se describen los requisitos de la herramienta y se expone un primer diseño del prototipo.
03/06/2021	Memoria y producto final.
10/06/2021	Presentación virtual.
Observaciones: En tratarse de un proyecto que incluye un marco teórico y otro práctico puede suponer movimientos en las fechas si no se dispone de toda la información necesaria para el desarrollo de la herramienta. Por otro lado, al crear una herramienta desde cero puede conllevar dificultades en el momento de la implementación del código.	

Tabla 1 Fechas claves y observaciones

1.7. Breve sumario de productos obtenidos

Durante el desarrollo del proyecto se obtienen una serie de productos a medida que se van completando las tareas descritas en el apartado anterior. Estos productos son los siguientes:

- La memoria del proyecto, siendo el presente documento donde se redacta todo el trabajo realizado en formato PDF.
- El código fuente desarrollado para la creación de la herramienta.
- El producto final, la herramienta de privacidad creada.
- Manual para instalar la herramienta en el lado del usuario.
- Las pruebas de usuarios para evaluar la viabilidad de la tecnología.
- La presentación virtual del proyecto.

1.8. Breve descripción de los otros capítulos de la memoria

Para el cumplimiento de los objetivos se considera realizar una primera parte de investigación en el área de tecnologías de mejora de la privacidad, por ello en la sección 2 se detallarán una serie de mecanismos de protección de privacidad que permitirán tener conocimiento de que técnicas se pueden desarrollar para llevar a cabo la implementación de la herramienta. Por otro lado, en la sección 3 se enumeran diversas tecnologías a nivel de usuario que otorgan una cierta privacidad en relación con la navegación y consultas web mostrando así la carencia de herramientas para este último campo. Para finalizar con la parte de investigación, en la sección 4 se puede encontrar WordNet, siendo la base de conocimiento seleccionada para la herramienta de privacidad. En este apartado, se encontrarán algunas de sus características principales y como se encuentra implementada.

La segunda parte de este proyecto está formado por la creación de la herramienta. En la sección 5, encontraremos los principales requisitos que debe cumplir un primer prototipo sobre el diseño, así como la implementación de una herramienta de consultas web mediante la generalización de consultas. Además, también se describirán los resultados obtenidos en las pruebas con usuarios reales junto con un breve análisis sobre una futura versión más completa.

Para concluir, en la sección 6 se exponen las conclusiones alcanzadas una vez finalizado el proyecto, en ellas se describe todo el proceso realizado, las dificultades encontradas durante el desarrollo al igual que todos los conocimientos adquiridos.

1.9. Terminología y conceptos clave

A pesar de la importancia que conlleva la seguridad informática hoy en día, la falta de conocimiento y desarrollo de herramientas de privacidad provoca que la terminología utilizada por este campo pueda ser difusa. Conceptos como privacidad, mecanismo o tecnología pueden tener diferentes significados según el contexto del problema que se quiera abordar.

En la literatura se puede encontrar diferentes significados para un mismo concepto, por este motivo se ve necesario definir una serie de términos que serán utilizados a lo largo del documento con la finalidad de dar una mayor comprensión al lector y no recurrir a la confusión.

En este proyecto se hablará de privacidad como una propiedad de la seguridad donde se establecen la confidencialidad y el control de los datos como prioritario. Por otro lado, se hace referencia a mecanismos de protección de datos y tecnologías de mejora de la privacidad. Seguidamente, se definirán los diversos significados que adquieren estas palabras y se indicará cuál de ellos se hace referencia en el documento.

El concepto de mecanismo dispone de diferentes significados, según las definiciones proporcionadas por el diccionario Merriam-Webster[1], un mecanismo es:

- Pieza de maquinaria.
- Proceso, técnica o sistema para conseguir un resultado.
- Operación o acción mecánica.
- Doctrina que sostiene que los procesos naturales (como de la vida) están determinados mecánicamente y pueden ser explicados por completo por las leyes de la física y la química.
- Procesos fundamentales involucrados o responsables de una acción, reacción u otro fenómeno natural.

En este proyecto, se hará uso de la segunda definición “Proceso, técnica o sistema para conseguir un resultado”. Concretamente como técnicas que se utilizan para implementar mejoras de privacidad en herramientas para el uso del usuario, teniendo como fin un mayor control sobre sus datos.

Respecto a las tecnologías de mejora de la privacidad o PET[2] se trata de un campo todavía desconocido para muchas personas. Este término se puede encontrar definido de varias formas, en [4] las definen como una variedad amplia de medios técnicos que tienen como fin proteger la privacidad de los usuarios. En cambio, se utiliza la denominación PET en el campo de las TIC que se encarga de proteger la privacidad, definiéndolo como un sistema coherente de medidas TIC que protegen la privacidad mediante la eliminación o reducción de los datos personales o previniendo el procesamiento innecesario y / o no deseado de datos personales [5].

En este proyecto, se utilizará el término de PET propuesto por ENISA [3] donde le da un significado más amplio definiéndolo como “sistemas que abarcan procesos, métodos o conocimientos técnicos para lograr una funcionalidad específica de privacidad o protección de datos o para proteger contra los riesgos a la privacidad de un individuo o grupo de personas físicas” [6].

Por lo tanto, el término establecido de tecnologías que mejoran la privacidad será utilizado para referirnos a un conjunto de herramientas y aplicaciones informáticas que implementan mecanismos de privacidad permitiendo al usuario proteger su información personal.

---

[1] https://www.merriam-webster.com/

[2] Privacy-enhancing technologies

[3] European Union Agency for Network and Information Security

1.10. Aspectos preliminares

El campo de la Web Semántica [7] surgió con el fin de definir la estructura del contenido que forma parte de la web. Su objetivo era crear un entorno donde las tecnologías que navegan por la red pudieran realizar tareas de reconocimiento de palabras clave que pertenecen a una página web.

Si bien su funcionamiento se ha llevado a cabo en muchas áreas, como es el caso del etiquetado colaborativo o los sistemas de recomendaciones, también puede ser utilizado para la realización de consultas web mediante buscadores semánticos. Estos buscadores utilizan la información semántica para mejorar la precisión de los resultados y proporcionan una respuesta más concreta a la consulta de un usuario. Un ejemplo de estos buscadores es DuckDuckGo.

En este proyecto, para crear la herramienta de protección de privacidad se pretende aplicar un mecanismo que tenga en cuenta la semántica de las consultas, de manera que se preserve el perfil de intereses ocultando aquella información personal de carácter sensible, pero procurando en todo lo posible mantener la semántica de las consultas originales. Para llevar esta implementación a cabo, se hará uso de las ontologías.

Las ontologías suelen estar escritas en lenguaje formal para poder ser interpretados por programas informáticos y tienen como fin facilitar la definición formal de conceptos representando categorías sobre un dominio determinado, así como la jerarquía que las engloba y las relaciones que las une entre ellas semánticamente.

Un ejemplo de uso es en los buscadores de contenido web, donde mediante las ontologías pueden realizar búsquedas que van más allá de una palabra clave, pudiendo buscar conceptos que a pesar de no aparecer literalmente en una página web, sí que aparece un significado similar o equivalente.

Actualmente podemos encontrar un gran número de ontologías que pueden ser utilizadas por los usuarios, una de estas tecnologías es WordNet, explicada en la sección 4. WordNet será utilizada con el fin de poner en práctica el mecanismo descrito en la literatura por [3], donde explica cómo mediante esta ontología podemos emparejar conceptos para recuperar nuevos relacionados semánticamente, ya sea bien hacia arriba o hacia debajo de la estructura del árbol. Estos conceptos serán utilizados para crear consultas que mantengan cierta relación con la consulta original.

Para facilitar la interpretación, los autores nos ponen el ejemplo de cómo este mecanismo funcionaria al realizar una consulta sobre “deportes acuáticos”, donde se podría generar una nueva consulta relacionada como “natación” (parte inferior del árbol denominada especialización) o bien con “deporte” (parte superior del árbol conocido como ancestro).

La implementación de este mecanismo se encuentra descrito en la sección 5.

2. Mecanismos de protección de privacidad

Una parte importante en el desarrollo de nuestra herramienta es conocer qué mecanismos existen en la actualidad para proteger nuestros datos y cuáles se ajustan mejor a nuestro objetivo. Estos mecanismos deben tener como fin proporcionar al usuario un mayor control de la información que desea compartir. Las técnicas propuestas en este apartado se podrán diferenciar en dos clases: técnicas criptográficas y técnicas perturbadoras de datos.

2.1. Mecanismos basados en la criptografía

Esta clase se encuentra formada por aquellos mecanismos capaces de llevar a cabo el anonimato. Comúnmente se le conoce como la ocultación de todos aquellos datos que puedan identificar a una persona, por lo que un atacante no puede identificar de forma clara al sujeto dentro de un conjunto.

Dentro de esta categoría existen varias modalidades, desde la aplicación de cifrados, ya sean bien públicos o privados, hasta algoritmos que desarrollan algún tipo de cifrado. La problemática que conlleva estos mecanismos es que el porcentaje de privacidad que ofrecen frente al servidor es bajo, ya que este último puede ser capaz de localizar al usuario mediante la dirección IP.

En relación con el anonimato existen varios mecanismos que tienen como fin proporcionar dicha ocultación. Un primer mecanismo conocido como k-anonimato viene derivado del concepto anterior, esta técnica intenta no revelar información que permita identificar a los individuos. Concretamente, su implementación se basa en considerar cada uno de los registros de un usuario como indistinguible al menos k-1 respecto a los registros del conjunto [8].

En cambio, existen otros mecanismos desarrollados a lo largo de los años por diferentes autores que implementan el anonimato mediante el cifrado de claves, como es el caso de Chaum [9]. Este autor desarrolló un mecanismo conocido como mixto(*mix*) que utiliza la recopilación de mensajes de la misma longitud de los remitentes y así, posteriormente, alterarlos criptográficamente mediante cifrado de clave privada y reenviarlos a sus destinatarios en un orden diferente.

Un mecanismo que está cogiendo fuerza en la actualidad es el cifrado homomórfico, es un método de cifrado que permite operaciones computacionales sobre datos cifrados. El mecanismo funciona generando un resultado cifrado, el cual una vez ha sido descifrado coincide con el resultado de las operaciones como si se hubieran realizado con datos no cifrados. Esta técnica otorga al propietario de los datos recibir los datos cifrados y así ver los originales sin ser alterados. En [10] proponen una arquitectura donde se emplea este mecanismo basado en un umbral de clave pública. El concepto de criptografía de umbral nos permite distribuir acciones de la clave privada entre los servidores, de manera que hasta que no colaboren, el texto cifrado no se puede descifrar.

Dentro de la criptografía, también existe otro mecanismo denominado computación segura de múltiples partes (SMPC[4]) siendo un subcampo del anterior. Este mecanismo tiene como fin preservar la privacidad de los participantes de una

[4] Secure Multi-Party Computation

conversación, donde utiliza la división de claves y así dificultar su intercepción en el proveedor del servicio.

Continuando con algoritmos criptográficos, existe una técnica denominada privacidad diferencial desarrollada para proteger a los usuarios en el momento de compartir información. Esta técnica agrega una capa de ruido al conjunto de datos que permite describir patrones de grupos dentro del conjunto mientras se mantiene la privacidad de las personas.

En la teoría también se han diseñado mecanismos basados en técnicas criptográficas. Se encuentra un ejemplo en el área de los servicios basados en la localización (LBS[5]) donde los autores llevan a cabo un mecanismo en la comunicación P2P[6] con un algoritmo de encubrimiento espacial para proteger la privacidad sin depender de una entidad externa centralizada [11]. El algoritmo permite formar parte de un grupo y calcula un área mínima que le permite cumplir con los requisitos de privacidad. Por el contrario, estos servicios pueden dar lugar a que los adversarios conozcan los hábitos e incluso rastreen las posiciones de un individuo.

Finalmente, se ha podido comprobar como los mecanismos criptográficos abarcan una gran variedad de posibilidades para otorgar anonimato, no obstante, dentro de esta clase también existen diversas técnicas que requieren la participación de usuarios externos. En este caso, no han sido consideradas puesto que para el objetivo final del proyecto no serían adecuadas, ya que tienen la desventaja de requerir un número elevado de usuarios si se quiere un tiempo aceptable en las consultas y al ser un factor que no se pueda predecir, provoca que la calidad del servicio se deteriore.

Por otro lado, en el supuesto de poder considerar en un futuro la participación de usuarios para proporcionar anonimato, un esquema a tener en cuenta sería el propuesto en [12], donde se considera un protocolo que utiliza herramientas criptográficas para desconocer la identidad del resto de usuarios, de modo que cada usuario enviaría una consulta de otro sin saber a quién pertenece.

2.2. Mecanismos de perturbación de datos

Una segunda clase de mecanismos son aquellos que tienen como finalidad perturbar los datos. El objetivo principal es proteger la información privada en situaciones donde el usuario comparte datos con terceros mediante la distorsión de su perfil. Una ventaja de utilizar técnicas de perturbación de datos es el hecho que los usuarios finales pueden adoptar en qué medida se realiza dicha perturbación, por tanto, les otorga mayor control y privacidad a sus datos personales.

En estos mecanismos se tiene en cuenta como ha sido construido el perfil del usuario, ya que la manera en que los datos estén representados permite utilizar un mecanismo u otro. Por ejemplo, en [13] modelan el perfil al clasificar la información del usuario en categorías permitiendo que estas sean simplificadas. Otro modelo de perfil es mediante etiquetas, donde daría un perfil más preciso, ya que son términos

[5] Location-based services

[6] Peer-to-peer

que han sido combinados con etiquetas semánticas enviadas por el usuario y serán utilizadas por el sistema para construir el perfil.

En los siguientes apartados se detallan diversas técnicas de perturbación que han sido descritas en la teoría con el fin de distorsionar los perfiles de usuarios, haciendo posible su aplicación en el campo de las consultas web.

2.2.1. Falsificación de datos

El primer mecanismo de perturbación encontrado es la falsificación de datos, donde se proporciona información que no refleja las preferencias reales. Un método utilizado para ello es agregar valores de forma aleatoria y luego enviar los datos perturbados a la otra parte pudiendo ser un sistema de recomendación, como por ejemplo un motor de búsqueda.

Esta técnica tiene varias desventajas, por un lado, al agregar valores aleatorios puede provocar que se den resultados alejados de los intereses del usuario, dando como resultado una búsqueda personalizada inservible. Por otro lado, en el caso de proporcionar intereses similares a los reales, a pesar de disfrazar los datos, estos pueden seguir revelando información sensible.

En el escenario de la Web Semántica, la falsificación de consultas consiste en acompañar consultas genuinas con falsas. En nuestro escenario el adversario sería el motor de búsqueda, el cual recibiría tanto consultas reales como falsas en nombre del usuario final, dando como resultado un perfil más impreciso. Este mecanismo garantiza la privacidad del usuario hasta cierto punto a costa del tráfico y los gastos generales de procesamiento.

El problema de la falsificación es que no preserva la veracidad de la información, por ello existen otros mecanismos que permiten un mejor ajuste del perfil de usuario respecto a sus intereses, como es el caso de la supresión de datos.

2.2.2. Supresión de datos

El mecanismo de supresión de datos es aquella técnica que permite al usuario abstenerse de proporcionar cierta información de forma que el perfil resultante de esta perturbación no capture sus intereses con tanta precisión.

En [13] proponen una arquitectura que implementa la supresión de etiquetas en la Web Semántica. Esta arquitectura ayuda a decidir al usuario qué etiquetas debe suprimir para obstaculizar a los atacantes y así impedir que construyan su perfil.

En la literatura [14] proponen utilizar este mecanismo en el campo del etiquetado colaborativo para proteger la privacidad ocultando cierto contenido. Los autores definen la supresión de etiquetas como una técnica que tiene como objetivo evitar que los atacantes de privacidad perfilen los intereses de los usuarios sobre la base de las etiquetas que especifican. Estos autores proponen una arquitectura construida sobre Delicious que permitirían a los usuarios indicar recursos de intereses donde se llevaría a cabo este mecanismo.

La supresión de datos se considera una técnica simple en términos de requisitos de infraestructura, no es necesario confiar los datos en una entidad externa. Por contra, tiene el costo de sobrecargar el procesamiento, así mismo la pérdida semántica conlleva una pérdida de integridad, una parte de la información nunca será enviada.

Una alternativa es la unión de los dos mecanismos anteriores, en [15] presentan un mecanismo de perturbación de datos en un sistema de recomendación donde los usuarios falsifican algunas calificaciones de elementos para no proporcionar un perfil tan preciso, además de poder abstenerse en aquellos elementos que ellos deseen.

Los autores lo nombran un mecanismo de calificación perturbadora pudiéndose llevar a la práctica mediante la implementación de un *software* instalado en la máquina local del usuario. Esta aplicación aconsejaría al usuario cuándo abstenerse de calificar un elemento dado y cuándo enviar calificaciones falsas para no reflejar su interés.

2.2.3. Generalización de datos

Por último, se encuentra el mecanismo denominado generalización de datos. La generalización transforma un conjunto de datos originales en un conjunto de datos falsificados mediante elementos más generales, es decir, reemplaza un elemento específico de carácter sensible, ya pueden ser etiquetas de usuario o en nuestro caso consultas, por categorías más genéricas de estas.

En el contexto de los sistemas de bases de datos relacionales, aplican estas técnicas para proporcionar anonimato generalizando y suprimiendo parte de los datos que se van a divulgar. En [16] presentan el concepto de generalización mínima, cuando los datos no se generalizan más de lo necesario para proporcionar anonimato, por ello, presentan un algoritmo para calcular el mínimo necesario que debe conllevar una generalización.

En el ámbito de consultas web, este mecanismo procura que se cree un perfil de usuario lo suficientemente preciso para obtener resultados útiles que se ajusten a sus intereses, pero sin poner en riesgo aquellos datos más sensibles. En la introducción de este proyecto hemos nombrado un par de esquemas que llevan a cabo la generalización de consultas mezclando consultas originales con consultas falsas basadas en este mecanismo. En [2] crean un protocolo para imitar el comportamiento de las consultas humanas gracias a la extracción de intereses encontrados en las redes sociales de los usuarios y así poder definir un perfil lo más ajustado posible.

La segunda propuesta explicada anteriormente utiliza un sistema configurable según la distancia semántica elegida por el usuario entre consultas originales y consultas creadas a través de la generalización [3]. Los autores utilizan este mecanismo para preservar el perfil manteniendo la semántica de las consultas originales en el momento de crear las nuevas, para ello analizan las consultas enviadas por el usuario interpretando la semántica de datos con la ayuda de bases de conocimiento ya construidas, marcándose como objetivo imitar el razonamiento humano.

La ventaja de este mecanismo es que no se llega a realizar una falsificación completa sino más bien una transformación, permitiendo mantener un mínimo los intereses del usuario. Este mismo hecho también lleva al problema de pérdida de información, por lo tanto, dará como resultado una consulta menos precisa que la original.

3. Tecnologías de mejora de la privacidad

En la actualidad muchos usuarios de internet todavía desconocen de qué herramientas disponen para proteger su información más confidencial.

En este apartado se van a enumerar una serie de tecnologías que tienen como fin proteger la privacidad de los usuarios, concretamente aquellas que puedan ser útiles en el área de consultas web.

Como se ha definido en el apartado 1.8, el término PET adopta muchos significados, en este proyecto adquiere la definición relacionada con aquellas tecnologías a nivel de usuario, es decir *software*, que están diseñadas con el fin de cumplir su objetivo sin poner en riesgo la privacidad y seguridad de las personas que hacen uso de él. En este aspecto, en [4] encontramos una clasificación de estas tecnologías según el objetivo que intenten abordar. Estos tres grupos son:

- Protección de la identidad, el principal objetivo es evitar revelar la identidad del usuario.
- Aislamiento, evitar que los usuarios sean molestados por contactos o solicitudes no deseadas, el más conocido es el correo *SPAM.*
- Control sobre los datos, permite que los usuarios puedan tener cierto control sobre sus datos personales, dando la posibilidad de recopilar, divulgar o determinar cómo utilizar sus datos o a quien son transferidos.

En este proyecto nos centraremos en las tecnologías que engloban el primero y tercer grupo, es decir, las tecnologías expuestas son las que implementan los mecanismos basados en el anonimato y técnicas de perturbación de datos.

Por otro lado, se hace una distinción de las tecnologías según el tipo de privacidad que presenten. Tal y como se explica en [17], las tecnologías expuestas se diferenciarán entre aquellas que aporten una privacidad *soft* o bien, aquellas que constan con una privacidad *hard*.

3.1. Tecnologías de privacidad soft

En esta categoría se encuentran aquellas tecnologías donde la protección de los datos viene dada por una entidad externa, los usuarios confían sus datos privados, por lo tanto, no habrá control de la información por su parte.

Una de las tecnologías más utilizadas para proporcionar privacidad en las búsquedas web a lo largo de los años son las tecnologías de comunicación anónimas (ACS[7]).

En la actualidad existen algunas de estas herramientas que proporcionan navegación web anónimas como es el caso de *anonymizer*, es un proxy web que elimina los encabezados de identificación y las direcciones de origen del navegador web. Actualmente existen varias aplicaciones gratuitas para el usuario, un ejemplo de este tipo de proxy es Kproxy [18], una extensión capaz de encriptar la conexión entre el navegador y el servidor de destino.

[7] Anonymous communication systems

A pesar de que son fáciles de utilizar, esta tecnología no evita el problema de privacidad, ya que únicamente se encargan de que el motor de búsqueda no perfile a los usuarios. En este caso, el proxy será quien creará el perfil lo que supone que un atacante puede conseguir dicha información atacándole directamente en lugar de al buscador.

Otra opción de comunicación anónima es el enrutamiento de cebolla [19] basado en el mecanismo de red mixta, donde su implementación la podemos encontrar en la herramienta TOR[8]. Esta herramienta es un sistema de enrutamiento de cebolla concreto que puede proporcionar comunicación anónima como navegación web, mensajería instantánea o aplicaciones que dependan de TCP[9].

La tecnología TOR [20] se basa en un conjunto de servidores donde permite a los usuarios conectarse a Internet mediante una serie de túneles virtuales en vez de realizar una conexión directa. Es una herramienta eficaz para evitar ser rastreados por sitios web, ya que utiliza criptografía de clave pública para encapsular el mensaje y asegurarse que únicamente el destinatario vea la información. En este caso se protege el transporte de datos, pero no es resistente a los ataques de análisis de tráfico siendo posible observar patrones de uso.

La desventaja de estas dos tecnologías es que son vulnerables frente a los atacantes pasivos que pueden llegar a monitorizar las comunicaciones que se encuentran en el mismo proxy. Puesto que solo se componen de un punto, recibir la respuesta a través de un canal anónimo consume mucho tiempo.

Por otro lado, los usuarios pueden decidir utilizar una serie de tecnologías más sencillas que les otorga un mínimo de anonimato como es el caso de I2P [21] o Private tunnel [22]. La primera tecnología consiste en una red privada gratuita a nivel mundial para comunicaciones seguras utilizando cifrado de extremo a extremo con el objetivo de ocultar el contenido de sus comunicaciones.

En el caso de Private tunnel, se trata de una VPN[10] que enmascara las direcciones IP públicas de los usuarios para que puedan navegar por la web de forma anónima y proteger sus redes contra ciberataques. Otra opción parecida a la anterior es NordVPN [23], herramienta que se utiliza para garantizar que toda información compartida a través de internet está encriptada y sea privada. Esta tecnología cifra los datos de un usuario antes de acceder a Internet a través de un túnel seguro.

Como se ha podido comprobar todas las tecnologías expuestas en este apartado dificultan que los datos privados de los usuarios sean robados, pero en ningún caso evitan que los buscadores construyan un perfil de usuario en base a ellos.

3.2. Tecnologías de privacidad hard

Las tecnologías que proporcionan este tipo de privacidad se caracterizan por no utilizar una entidad externa, puesto que se considera que el usuario es consciente de la importancia de proteger sus datos y se encarga él mismo de protegerlos.

En este apartado se encuentran las tecnologías que implementan mecanismos de perturbación de datos. En la sección 2 se han nombrado varias técnicas que podrían

[8] The Onion Router

[9] Transmission Control Protocol

[10] Virtual Private Network

llegar a implementarse en futuras herramientas para la mejora de privacidad de consultas web. No obstante, en la actualidad este tipo de aplicaciones es escaso y únicamente existen dos opciones a nivel de usuario disponibles: TrackMeNot y GooPIR.

TrackMeNot [24] [25] es una herramienta que tiene como fin proteger la privacidad en las consultas web mediante el mecanismo basado en la falsificación de consultas. TMN se encuentra disponible en forma de complemento para los navegadores Firefox y Google y ha sido implementado en JavaScript, C++ y XUL. Su funcionamiento se basa en la generación dinámica de consultas falsas mezcladas con aquellas consultas reales que el propio usuario envía al motor de búsqueda. Estas consultas aleatorias se crean gracias a una recopilación de frases similares a consultas web que han sido realizadas recientemente y se envían a través de solicitudes HTTP a los motores de búsqueda que el usuario especifica.

En el caso de GooPIR [26] [27], es un prototipo desarrollado en java que implementa el mecanismo de falsificación de consultas, pero a diferencia del anterior oculta la consulta real mediante un conjunto de palabras falsas.

Cuando un usuario envía una consulta, la aplicación obtiene palabras clave con la misma frecuencia de uso que tiene la consulta del usuario, estas palabras se adjuntan a la consulta con el fin de confundir al motor. Una vez obtenido los resultados de la búsqueda, GooPIR devuelve al usuario aquellos que sean relevantes con su consulta inicial.

Así como TMN utilizaba consultas recientes para realizar la falsificación, GooPIR utiliza cualquier tesauro de referencia para decidir qué conceptos agregar, por ello únicamente puede enviar palabras.

Estas dos tecnologías tienen la ventaja de distorsionar el perfil de usuario al enviar consultas aleatorias al motor de búsqueda. Sin embargo, el envío dinámico de consultas falsas provoca un aumento en el tráfico de la red y sobrecarga los motores de búsqueda reduciendo así su rendimiento. Además, pueden ser vulnerables a ataques de análisis semántico sobre consultas agregadas que podrían revelar intereses reales de los usuarios.

En el caso de TMN se encuentra una segunda desventaja respecto al envío de consultas cuando el usuario no se encuentra presente debido a que el motor de búsqueda puede analizar los patrones y dividir el perfil según el ritmo de consultas enviadas, siendo posible deducir si una determinada consulta fue enviada por la herramienta. Un ejemplo de ello se puede encontrar en [28], donde los autores evidencian que es posible distinguir consultas reales de consultas creadas por esta herramienta.

3.3. Recomendaciones para mejorar la privacidad del usuario

Como se ha podido observar no existen tecnologías disponibles a nivel de usuario que tengan como fin el objetivo de este proyecto, una herramienta basada en la generalización de consultas. Por ese motivo, se considera apropiado nombrar algunas recomendaciones que el usuario puede adoptar de forma sencilla en su navegador de confianza para otorgarle un extra en su privacidad.

La primera recomendación es utilizar direcciones de IP dinámicas con un navegador web sin cookies. Para ello se puede utilizar la pestaña de incógnito que navegadores como Firefox o Google Chrome proporcionan, ya que permiten disminuir la huella

digital del usuario eliminando las cookies y los archivos temporales una vez se cierra el navegador. Si no se desea utilizar dicha pestaña, una alternativa es desactivar manualmente el historial web y así eliminar todas las páginas visitadas en el momento que salga del navegador.

El problema de este mecanismo es que solo oculta la identidad real del usuario, pero las consultas siguen siendo las mismas. Por otro lado, la política de renovación de IP viene dada por el operador de red y este puede dar siempre la misma.

Siguiendo con la navegación, es conveniente disponer de un cortafuegos que permita supervisar y filtrar el tráfico de red mediante políticas, donde el usuario puede especificar qué conexiones son confiables y así comunicarse con el sistema.

En relación con las consultas web, existen varios motores de búsquedas no tan populares como StartPage o Gibiru que están diseñados para realizar búsquedas privadas. StartPage [29] se caracteriza por no registrar el historial de búsqueda, evitar el almacenamiento de cookies y los anuncios personalizados. En el caso de Gibiru [30] asegura no recopilar datos del usuario ni almacenar la dirección IP.

Por último, en ser un proyecto que desarrolla una tecnología de mejora de la privacidad sobre el buscador de Google, es conveniente que los usuarios tengan conocimientos sobre qué políticas de privacidad sigue dicho motor acerca de los datos que recolectan y cómo son gestionados, estas políticas pueden ser encontradas en el centro de políticas de Google[11].

En el caso de las búsquedas de Google, tanto si el usuario ha iniciado sesión como si se hace un uso esporádico, el buscador utiliza cookies para mejorar su servicio y mostrar anuncios y contenidos personalizados. Esta configuración se encuentra por defecto, por lo que si el usuario no desea una búsqueda personalizada es necesario cambiar la configuración. Si se realizan las búsquedas desde una pestaña privada, el propio buscador te muestra un recordatorio antes de realizar la búsqueda donde permite personalizar el uso de las cookies. En ese caso, el usuario únicamente debe marcar “No” a todas las opciones que desee y posteriormente confirmar.

Ilustración 3 Ajustes de personalización de Google

[11] https://policies.google.com/

4. WordNet

Durante el desarrollo de este proyecto se utilizará como base de conocimiento WordNet [31]. Para poder hacer un buen uso de esta herramienta es necesario conocer de qué se trata, así como qué elementos la componen y su implementación. En los siguientes apartados se explicará de manera detallada en qué consiste esta tecnología junto con cada una de las categorías y relaciones que podemos encontrar en ella.

4.1. WordNet – ¿En qué consiste?

Es una base de datos léxica escrita en inglés, donde se pueden encontrar sustantivos, verbos, adjetivos y adverbios agrupados en conjuntos de sinónimos conocidos como *synsets*, haciendo un total de 117.000 en la base de datos. Cada uno de ellos contiene una definición breve conocida como *gloss*, donde detalla el uso del conjunto.

Esta base de datos se encuentra pública y gratuita, tanto para el navegador como para descargar. Es una herramienta útil para la lingüística computacional y el procesamiento del lenguaje natural.

4.2. Subredes

WordNet [32] consta de 4 subredes: sustantivos, adjetivos, adverbios y verbos.

En cada una de ellas se diferencian dos conceptos: las formas de palabra y el significado de palabra. El primer término se utiliza para referirse a la expresión física de la palabra. El segundo se trata del concepto léxico para expresar lo anterior. Algunas formas pueden tener diferentes significados y algunos significados pueden ser expresados de diferentes formas. Aquellas formas de palabras que puedan obtener diferentes significados serán representadas por synsets distintos, siendo cada uno de ellos únicos.

Los sustantivos y los verbos se organizan mediante jerarquías, pudiendo tener punteros cruzados de referencia entre ellos. Además, los verbos se encuentran divididos según la acción o el evento que representen, aquellos verbos que se encuentren en la parte inferior de la jerarquía son más característicos respecto a un evento.

En relación con los adjetivos, WordNet los divide en dos clases: descriptivos y relacionales. Los adjetivos descriptivos se caracterizan por expresar el valor de un sustantivo. En cambio, los relacionales son aquellos que se crean a partir de un sustantivo.

4.3. Relaciones semánticas y léxicas

WordNet distingue entre relaciones semánticas y léxicas según la subred que se esté tratando. A continuación, se detallan las más utilizadas para cada una de las subredes anteriormente nombradas.

En el caso de los sustantivos, la relación más frecuente entre los synsets es la hiponimia. Consiste en una relación semántica entre significados, es transitiva y asimétrica. Un hipónimo hereda todas las características del concepto más genérico conocido como hiperónimo, y agrega al menos una característica que lo distingue de su superior y de cualquier otro de esa jerarquía.

Otra relación es la meronimia (parte - todo), también es transitiva y asimétrica y se mantiene entre synsets donde las partes se heredan de sus superiores. En el caso de los verbos, existe también una relación particular llamada toponimia, la cual expresa que un verbo muestra cierto vínculo con otro verbo más general.

Para los adjetivos se puede diferenciar relaciones según la clase. En el caso de un adjetivo descriptivo, la relación más utilizada es la antonimia, los pares de antónimos directos que a su vez están vinculados a varios semánticamente similares, forman también la relación de antónimos indirectos respecto al polo opuesto. Para aquellos adjetivos que son relacionales se encuentra vinculados con los sustantivos de los que se derivan. Por otro lado, los adverbios se encuentran vinculados a los adjetivos, puesto que en inglés derivan de ellos.

En WordNet también se incluyen vínculos morfosemánticos que se mantienen entre palabras semánticamente similares que comparten una raíz con el mismo significado. En los pares sustantivo - verbo se especifica el papel semántico del sustantivo respecto al verbo.

4.4. Implementación

La base de datos se encuentra compuesta por dos ficheros para cada una de las categorías, *index* y *data*. Ambos ficheros se encuentran interrelacionados e implementados en formato ASCII. El fichero *index* detalla una lista de todas las formas de palabras que incorpora WordNet. Mientras que, *data* contiene los datos de todos los archivos de entrada de la base de datos.

En cuanto a las interfaces de usuario, se basan en una biblioteca de funciones que interactúan con el segundo fichero. Las funciones encontradas permiten manipular las cadenas de búsqueda de manera simple, únicamente se debe especificar la palabra deseada en el buscador y WordNet mostrará todos los significados que esta palabra adquiere. Seguidamente, tal y como se muestra en la figura, se facilitan las diferentes relaciones explicadas mediante un recuadro donde el usuario puede seleccionar cuales desea ver.

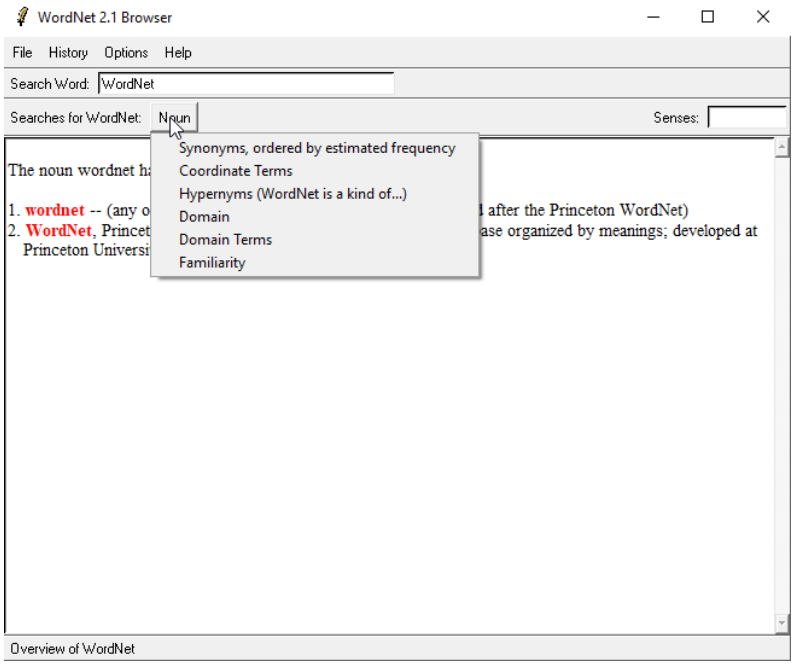


Ilustración 4 Herramienta WordNet para escritorio

5. Desarrollo de la herramienta de privacidad

En esta sección se describe y desarrolla la primera versión de una herramienta de privacidad, marcando como objetivo brindar a los usuarios un mayor control sobre sus datos personales gracias a la implementación de un mecanismo de distorsión de consultas web, obligando al motor de búsqueda de Google crear un perfil con intereses más generales y, por lo tanto, menos sensibles.

Este desarrollo está dividido en diferentes fases que corresponden a cada uno de los subapartados siguientes. Cabe recalcar que este proceso se ha realizado de forma iterativa, pero únicamente se muestra el resultado final de cada una de las cuatro etapas. No obstante, se irá remarcando qué aspectos se han ido mejorando a lo largo de cada una de ellas.

Ilustración 5 Metodología empleada en el desarrollo de la herramienta

La primera fase descrita corresponde a la funcionalidad principal de la herramienta, así como una serie de requisitos que debe incorporar para que cumpla con la meta marcada. En el apartado 5.2 se lleva a cabo una segunda fase tratándose del diseño de la herramienta, donde se podrá observar un primer prototipo creado a partir de los requisitos especificados. Seguidamente, en el apartado 5.3 da lugar al desarrollo de la aplicación en el que se mostrará los diferentes códigos implementados para que la herramienta cumpla con sus funciones. A pesar de tener ambas fases separadas, tanto la creación del diseño como el desarrollo de esta se ejecutan a la vez, de modo que se van realizando cambios según aparecen nuevas situaciones.

Una vez se tenga una herramienta funcional es importante empezar hacer pruebas con ella para ir analizando su viabilidad y rendimiento. Para ello, se realizan pruebas con usuarios reales, en la sección 5.4 se describen dichas pruebas junto con las conclusiones extraídas en el proceso.

Por último, se encuentra un apartado donde se hablará sobre una posible segunda versión que la herramienta pueda conllevar.

5.1. Definición y requisitos

La herramienta propuesta tiene como finalidad mejorar la privacidad del usuario frente a la recolección de datos personales que realizan los motores de búsqueda en el momento de realizar consultas web.

El prototipo por desarrollar debe poseer la funcionalidad principal de generalización de consultas, donde a partir de una consulta específica se busca un concepto relacionado semánticamente más general. Esta funcionalidad será descrita más detalladamente en la sección 5.3.

En un principio se había planteado únicamente la opción de disponer de consultas privadas, donde se llevaría a cabo la generalización y por lo tanto una consulta menos ajustada a sus intereses. Sin embargo, se plantea la opción de que sea operativa para cualquier tipo de consulta sin tener que recurrir al motor de búsqueda, por esta razón y para seguir aportando al usuario cierta protección respecto a sus datos, se decide crear un tipo de consulta personalizada, donde se mantendrá una relación semántica más estrecha.

Por otro lado, a partir del estudio previo realizado, se han encontrado una serie de requisitos funcionales que nuestra herramienta debería implementar. Todos los expuestos en la siguiente tabla son esenciales, por lo que su prioridad durante el desarrollo de la herramienta es alta y deben cumplirse.

Requisito	**Descripción**	**Restricciones**
R01	Proporcionar consultas para los diferentes niveles de complejidad. Desde consultas simples, formadas por una palabra, hasta consultas complejas, tratándose de un conjunto de palabras o frases.	El idioma de las consultas deberá ser en inglés. Para las consultas complejas únicamente se tendrán en cuenta frases nominales.
R02	Definir el tipo de consulta. Si el usuario desea una búsqueda personalizada, debe tener una mayor vinculación con la consulta, por lo tanto, una relación semántica más cercana. En cambio, si prefiere que el dato no sea conocido deberá seleccionar una búsqueda privada.	La herramienta solo contempla la generalización descrita por los dos tipos de consulta.
R03	Disponible para Google Chrome. Al principio del presente documento se presentaron unas estadísticas del último año donde colocaban a Google como el mayor motor de búsqueda utilizado, por ello es conveniente que esta herramienta tenga como requisito ser funcional para dicho buscador.	La herramienta se deberá instalar en el propio navegador para ser utilizada.

R04	Funcional desde cualquier ventana. Permitir al usuario utilizar la herramienta en cualquier momento mientras navega por otras páginas webs sin tener que dirigirse al buscador.	
R05	Conexión al servidor. La herramienta necesita un servidor donde alojar la función que consulta la base de conocimiento utilizada para la generalización, en este caso WordNet.	El servidor solamente dispondrá de peticiones *GET*.

Tabla 2 Requerimientos de la herramienta

5.2. Diseño

En este apartado se exponen los criterios de diseño adoptados para la interfaz de usuario, además se muestra un primer prototipo de la herramienta. En este caso, la fase de diseño parte de los requisitos anteriormente expuestos y se encuentra estrechamente unida al proceso de desarrollo. Este hecho conlleva que el diseño de la interfaz vaya variando según se aplican o modifican las diferentes funcionalidades requeridas.

El proceso comenzó teniendo en consideración los requisitos R01 y R02, ya que son las características principales que el prototipo debía de contemplar. En primer lugar, teniendo en cuenta lo indicado en R01, el diseño debe incluir un buscador donde el usuario pueda realizar la búsqueda específica que desee, es decir, un elemento donde escribir las consultas simples o complejas. Para el requisito R02, es necesario un parámetro configurable donde el usuario pueda elegir el tipo de consulta que desee, siendo posible escoger entre una consulta personalizada o bien una consulta privada.

Por otro lado, para que sea una herramienta fácil e intuitiva de utilizar y cumpla con el requisito R03, se toma la decisión de diseñarla como una extensión para el navegador web Google Chrome, donde una vez instalado se pueda hacer uso de ella mediante el icono de la extensión que se encontrará en el margen superior derecho dentro de la barra de navegación, cumpliendo así con R04. En la siguiente ilustración se muestra el prototipo final sin estilos de diseño, pero visibilizando las funcionalidades descritas.

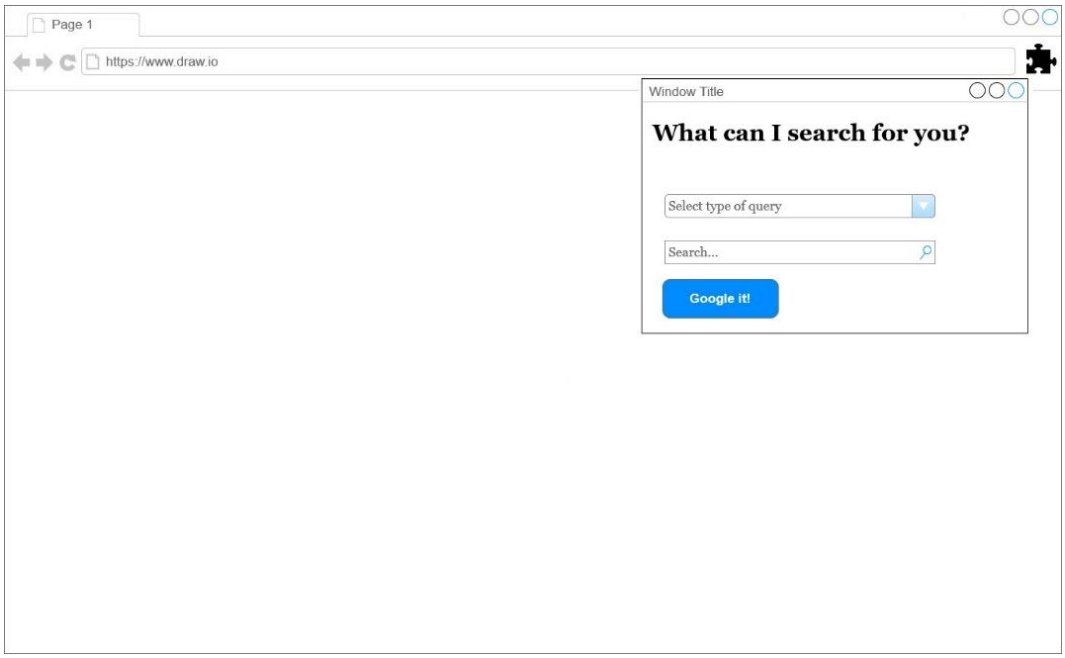


Ilustración 6 Wireframe de la extensión

Como se puede observar, el usuario podrá interaccionar con ella mediante un campo para realizar la consulta, en él deberá escribir la misma búsqueda que realizaría si fuera Google. El sistema también incluye un desplegable para poder identificar el tipo de consulta, este componente tendrá dos opciones disponibles: personalizada o privada. Una vez el usuario haya indicado los dos valores anteriores, únicamente deberá hacer clic en el botón para enviar la consulta o bien hacer clic en la tecla *Enter* del teclado.

5.3. Desarrollo

La fase de desarrollo, también conocida como fase de implementación, es la etapa más extensa de todas, donde se crea tanto la interfaz de usuario como las funcionalidades descritas, dando como producto final la extensión de navegador.

Durante el desarrollo se tiene en cuenta las dos fases previas, por esta razón, a lo largo del apartado se describirán diferentes obstáculos que han hecho cambiar aspectos del diseño, pero como se ha explicado anteriormente, solo se muestran los resultados finales obtenidos.

Antes de empezar a explicar cada una de las implementaciones realizadas, es conveniente conocer qué tecnología se está llevando a cabo. La herramienta de privacidad ha sido desarrollada como una extensión para Google Chrome, siendo un pequeño programa que se instala dentro del navegador y añade funciones al mismo. Actualmente, las extensiones están disponibles para cualquier navegador moderno, siendo Google Chrome quien ofrece una cartelera más extensa, desde extensiones para proteger al usuario de programas malignos hasta bloqueadores de publicidad.

El proceso de creación de una extensión es muy sencillo de realizar, en el caso de Chrome se ofrece una API para dar una mayor facilidad al desarrollo. La implementación se realiza como el *frontend* de una página web, en el apartado 5.3.2 se profundiza sobre ello, enumerando los diferentes archivos requeridos para su funcionamiento, así como el proceso seguido para su creación.

En el caso de nuestra herramienta, a parte de la extensión que se encontrará en el lado del usuario, también se requiere un lado del servidor para su correcto funcionamiento, ya que será el encargado de almacenar la función para generalizar las consultas.

A continuación, se nombran todas las tecnologías utilizadas, así como el lenguaje de programación implementado para poder llevar a cabo todo el proceso de desarrollo. Más adelante, se describirán los pasos seguidos para crear la extensión del navegador Chrome y la función para generalizar la consulta. Por último, se explicará cómo se realiza la conexión entre el lado del usuario y el servidor.

5.3.1. Software

Para poder desarrollar la herramienta es necesario hacer uso de las siguientes tecnologías:

- Editor de código.
- Lenguajes de programación: Python, JavaScript, HTML y CSS.
- Framework Flask.
- Librería NLTK incluida en Python que incorpora la base de datos WordNet.
- Servidor para el generalizador de consultas (Gunicorn).
- Plataforma de servicios en la nube (Heroku).

5.3.2. Extensión Google

Las extensiones de Google Chrome [33] se componen por tres ficheros: el manifiesto, la interfaz de usuario y el script de contenido. En este caso, también se ha incluido un archivo para el estilo gráfico de la interfaz, siendo este último totalmente opcional.

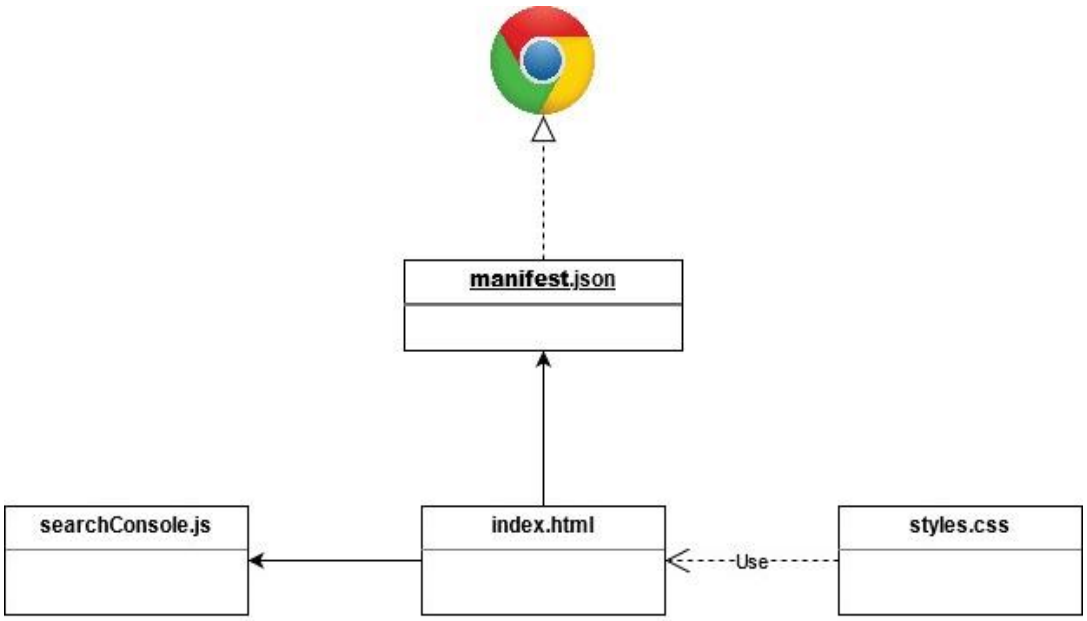


Ilustración 7 Componentes de una extensión de Google Chrome

A continuación, se explican los principales pasos para crear la extensión diseñada en el apartado anterior:

i. Se crea una carpeta donde almacenar los diferentes archivos que requiere la extensión. Esta carpeta debe nombrarse del mismo modo que lo hace nuestra herramienta. En este caso, se ha decidido poner el nombre de JCM.

ii. Una vez disponemos de la carpeta, el primer archivo que toda extensión debe incluir es el manifiesto, en la ilustración 6 lo encontramos con el nombre *manifest,* siendo un archivo en formato JSON. Este archivo es el más importante, ya que se encarga de proporcionar al navegador toda la información acerca de la extensión y así proceder a su uso.

   Un ejemplo de que contenido debe incluir se muestra en la siguiente imagen, en este caso corresponde al manifiesto creado para este proyecto. En él se especifican los campos correspondientes al nombre y versión de la extensión, la versión del manifiesto junto con las acciones y permisos requeridos.

Ilustración 8 Archivo manifiest.json

iii. El segundo archivo es el indicado con el nombre index.html. En este archivo se encontrará todo el código relacionado con la interfaz del usuario, por lo tanto, corresponde a un archivo HTML donde además se añade la llamada al script de contenido.

   Tal y como se pudo ver en el apartado de diseño, se ha optado por crear una ventana emergente para interaccionar con el usuario, la cual se muestra una vez el usuario hace clic en el icono de la parte superior derecha del navegador. Para que esto sea posible, dentro del manifiesto se debe asignar la acción *default_popup* a este archivo.

iv. El tercer archivo creado es el correspondiente a searchConsole escrito en JavaScript y llamado por el archivo HTML. Este archivo es conocido como el script de contenido, siendo el responsable de la conexión entre la extensión y el servidor para posteriormente enviar la consulta a la página web de Google.

```
async function handleQuerySubmit(event) {
  event.preventDefault();

  var request = new XMLHttpRequest();
  let query = input.value;
  let typeq = type.value;

  q = `query=${query}&type=${typeq}`;
  request.open("GET", "https://jcmserver.herokuapp.com/getmsg/?" +q , true);

  request.onreadystatechange = function()
  {
    if (request.readyState == 4 && request.status == 200)
    {
      var result =  JSON.parse(request.responseText);
      var newURL = 'https://www.google.com/search?q=' + encodeURIComponent(result);
      chrome.tabs.create({ url: newURL });
    }
  }
  request.send();
}
```

Ilustración 9 Código searchConsole.js

En la ilustración 8 se muestra parte del código desarrollado para realizar la conexión. Los pasos de la implementación son los siguientes:

1- La función captura la consulta que el usuario ha indicado previamente en el buscador junto con el tipo de consulta que desea realizar.

2- Una vez obtenido los valores, se unen ambas respuestas para poder pasar el resultado a través de una petición HTTP ya que, para obtener la respuesta del servidor, es necesario pasarle los parámetros por la URL.

3- Se realiza la petición HTTP mediante el método GET, además es necesario incluir la URL, los parámetros anteriores y un tercer parámetro con el valor true para indicar que la petición sea gestionada de forma asíncrona y así evitar reanudar la ejecución. La conexión entre esta función y el servidor será detallada en los próximos apartados.

4- Una vez se haya realizado la generalización, se le devuelve la nueva consulta que será enviada directamente a Google. Para que esto sea posible, se hace uso del evento *onreadystatechange()* donde indicamos que se debe enviar la respuesta obtenida al buscador de Google cada vez que se cambia de estado.

Para finalizar, he de recalcar que esta función genera una URL nueva especificando la búsqueda en Google que se quiera realizar, por tanto, se abrirá una nueva pestaña cada vez que se mande una consulta. Por otro lado, la generalización únicamente se efectúa en la ventana emergente, esto significa que si el usuario se dirige al buscador para realizar la consulta no se hará ningún tipo de modificación en ella.

v. El último paso del proceso es la creación de un archivo CSS. Su función es proporcionar al HTML los diferentes estilos gráficos de la extensión.

5.3.3. Generalizador de consultas

Una vez conocemos como se ha desarrollado la parte relativa a la interfaz damos paso al generalizador de consultas, siendo una función implementada en Python.

En un principio se consideró implementar todo tipo de frases complejas, pero a causa de la estructura interna de WordNet, resultaba ser un código más engorroso de ejecutar si se pretendía diferenciar entre sustantivos y verbos. Puesto que el objetivo principal de este proyecto no necesitaba tal nivel de profundidad, finalmente se optó por una función donde únicamente contemplase las consultas simples y complejas no verbales.

```
def respond():
    # Retrieve the query and type from url parameter
    query = request.args.get('query')
    qtype = request.args.get('type')
    newQueryList = []
    wordList = query.split()

    for word in wordList:
        try:
            w = wn.synsets(word)[0].name()
            syn = wn.synset(w)

            if qtype == 'private':
                hypernym = syn.hypernyms()[0]
                newQueryList.append(hypernym.lemma_names()[0].replace("_"," "))

            elif qtype == 'personalized':
                if len(syn.lemma_names())>1:
                        newQueryList.append(syn.lemma_names()[1].replace("_"," "))
                else:
                        newQueryList.append(word)
            else:
                newQueryList.append(word)
        except:
            newQueryList.append(word)

    newQuery = ' '.join(newQueryList).capitalize()
    response = newQuery
```

Ilustración 10 Código para generalizar consultas

El código que lleva a cabo la generalización de la consulta se encuentra en *respond()*. Esta función recoge los parámetros indicados en la URL correspondiente a la consulta que el usuario ha enviado al buscador y el tipo de consulta que desea.

Respecto al primer parámetro, hay que tener en cuenta que el algoritmo no es capaz de identificar el significado propio que tiene una palabra dentro de una frase, por ello cuanto más compleja sea la consulta, mayor desvinculación habrá respecto a la original, siendo posible no llegar a realizar una búsqueda adecuada con los intereses del usuario.

En relación con el segundo parámetro, solo habrá dos valores posibles: privado o personalizado. Según el tipo de consulta especificado, la función dará una respuesta semánticamente más similar a su palabra de origen.

- Si la consulta realizada es privada, se requiere obtener una palabra más alejada de la consulta original. Por este motivo, la función buscará el hiperónimo más cercano de la palabra indicada, es decir, devolverá la primera palabra de la rama superior del árbol.
- En el caso de una consulta personalizada, se debe obtener una consulta más cercana a la original, devolviendo un sinónimo de la palabra. Si se da el supuesto de no disponer de ninguna, entonces se devolverá la palabra original.
- También se considera la situación de que el usuario realice la consulta sin indicar ningún tipo de privacidad, en este caso retornará la primera consulta.

Por último, durante el desarrollo del código se han debido tomar una serie de decisiones para poder tener la mejor funcionalidad posible:

- Para poder hacer uso del generalizador en consultas complejas, primero de todo se deben separar cada una de las palabras y realizar la generalización por separado. Una vez se hayan obtenido los resultados, se vuelve a construir la frase en el mismo orden que la original.
- En el momento de buscar la palabra en WordNet, se toma la decisión de seleccionar el primer *synset* que corresponde al significado común de dicha palabra.
- En el caso que la palabra no se encuentre en WordNet, la función devolverá la original.
- La elección del hiperónimo seguirá la misma regla que en el caso anterior, se devolverá aquel que se encuentre en la primera posición y así otorgar una mínima vinculación con la consulta original.
- Para los sinónimos, si la palabra original dispone de más de uno se devolverá el que se encuentra en la segunda posición, puesto que WordNet cuenta la palabra original como el primer sinónimo dentro del *synset*.

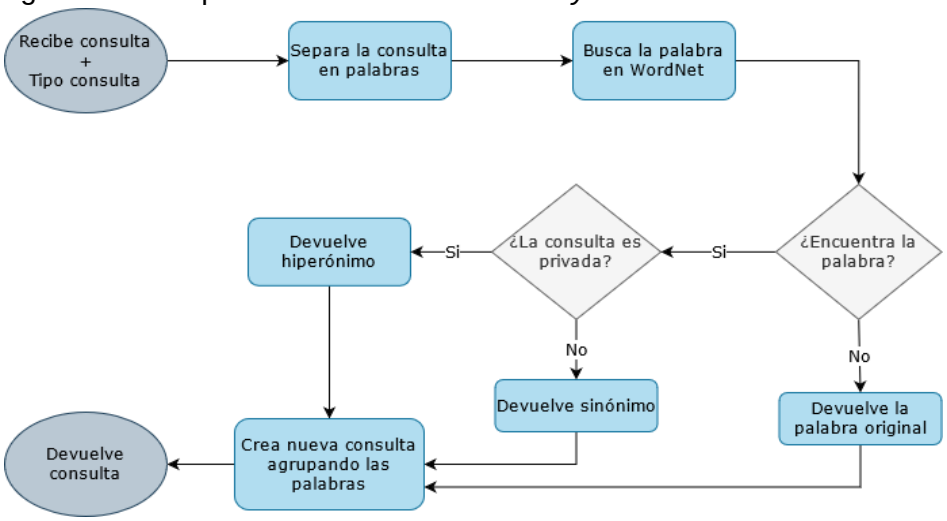


Ilustración 11 Diagrama de flujo Generalización de consultas

5.3.4. Infraestructura usuario - servidor

Para poder poner en funcionamiento la extensión es imprescindible una conexión entre la interfaz de usuario y el generalizador de consultas.

Como se ha podido ver a lo largo de la sección, el desarrollo de ambas partes se ha realizado en lenguajes diferentes, en consecuencia, se requiere de un mecanismo para poder compartir los datos. A continuación, se describirá la creación del servidor que almacena el código ilustrado en el apartado anterior, permitiendo la conexión con la extensión, concretamente con el fichero de JavaScript.

El lado del servidor conocido como *backend* se encuentra formado por dos elementos: el código de la aplicación para realizar la conexión al servidor donde se almacenará el generalizador de consultas y un servicio de computación en la nube.

Para la conexión al servidor se ha tomado la decisión de hacer uso de Flask con el servidor HTTP[12] Gunicorn. Flask es un *framework* escrito en Python que ejecuta de forma rápida aplicaciones webs básicas y conexiones a servidores. Además, permite definir funciones que corren una vez solicitada una ruta específica junto con la petición requerida que conlleva dicha ruta.

Por otro lado, tenemos el servidor HTTP, en este caso se ha escogido Gunicorn, siendo un servidor HTTP específico para Python que cumple con la especificación WSGI[13]. Este servidor permite trabajar con la aplicación Flask, ya que administra las peticiones simultáneas que la aplicación recibe.

Una vez se ha creado el servidor es necesario alojarlo en alguna plataforma para poder realizar la conexión desde cualquier host de Internet. Dentro del mercado de plataformas de computación en la nube, existen numerosas posibilidades para almacenar código y llevar a cabo el despliegue de una aplicación, como es el caso de AWS, Google Cloud, Digital Ocean o Heroku, siendo esta última la seleccionada para almacenar el servidor.

Heroku [34] es una plataforma de servicios en la nube que permite alojar aplicaciones y manejar los servidores de forma sencilla, únicamente hacen falta cuatro ficheros para el despliegue del servidor.

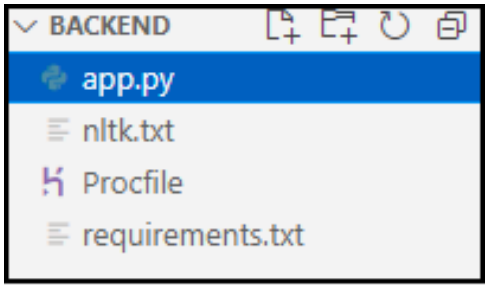


Ilustración 12 Esquema servidor Heroku

El esquema mostrado en la imagen corresponde al directorio raíz de la aplicación ubicado en el servidor.

[12] Hypertext Transfer Protocol

[13] Web Server Gateway Interface

Cada uno de estos archivos cumple una funcionalidad específica:

- App.py, código Python que lleva implementada la función descrita en el apartado 5.3.3.
- Nltk.txt, archivo específico para la librería que incluye la base de datos de WordNet.
- Procfile, corresponde a los comandos que deben ser ejecutados dentro del contenedor, en este caso únicamente hace falta el comando *web: gunicorn app:app.*
- Requirements.txt, archivo de texto que incluye todas las versiones de los diferentes paquetes que deben ser instalados para que el despliegue funcione correctamente.

Esta plataforma es considerada una de las más simples de desplegar. Para poder ejecutar el servidor en Heroku se siguen los siguientes pasos:

1- Se crea el directorio anterior donde se llevará a cabo el servidor.

2- Para inicializar el proyecto, se utiliza un control remoto mediante el comando *heroku create*.

3- Una vez tenemos nuestro proyecto, se crea el procfile que será utilizado cuando comience Gunicorn, es decir, se inicia el servidor.

4- El último archivo por crear será requirements.txt.

5- En el caso de realizar cambios y reiniciar la aplicación, Heroku utiliza comandas Git para facilitar el proceso, del mismo modo que se realizaría en un repositorio de este servicio.

Una vez hemos alojado el directorio en nuestra plataforma y el servidor se encuentre activo, entonces se podrá realizar la conexión mediante la URL indicada:

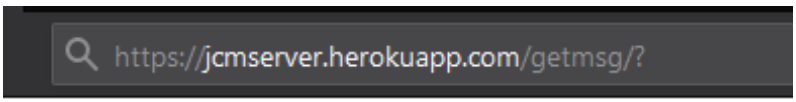


Ilustración 13 URL para la conexión extensión - servidor

Como se ha podido observar se trata de una plataforma fácil que permite alojar todo tipo de aplicación web con solamente cinco pasos. Además, otro punto fuerte que ha influenciado en su elección es la disponibilidad de una versión gratuita para proyectos pequeños como es el caso de la extensión, permitiendo ejecutar las pruebas con usuarios desde sus respectivos equipos personales.

Como desventaja se podría decir que, al tratarse de una versión gratuita, el servidor entra en reposo cada 30 minutos sin tráfico, este factor se debe tener en cuenta en el momento de realizar una primera conexión, ya que puede demorar los resultados de las primeras consultas.

Para finalizar, se presenta un esquema de la infraestructura completa incluyendo cada uno de los componentes que forman parte de la conexión entre el lado del usuario y del servidor.

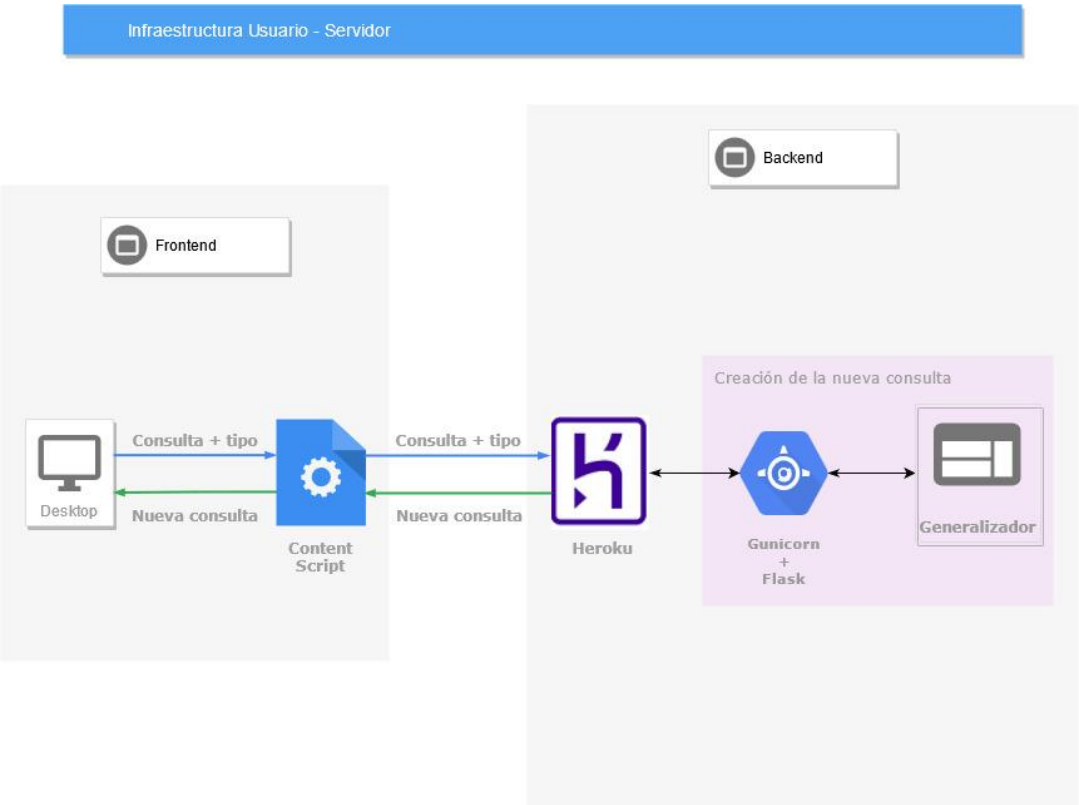


Ilustración 14 Infraestructura Usuario - Servidor

5.4. Evaluación

En este apartado se exponen las pruebas realizadas con usuarios reales, así como los diferentes resultados obtenidos. El objetivo es conocer qué opiniones tienen los usuarios respecto a las consultas obtenidas por nuestra herramienta.

Antes de empezar con las pruebas se les explica a los participantes que las consultas serán en inglés, así como una breve explicación sobre el funcionamiento de la extensión y las principales diferencias entre los tipos de consultas disponibles.

Una vez el usuario dispone de toda la información necesaria, se da comienzo a la misma realizando un par de preguntas en relación con la privacidad y las extensiones. Concretamente se plantean dos preguntas:

- ¿Toma alguna medida para proteger sus datos privados? En caso afirmativo, podría decirme cuál.
- ¿Sabría instalar una extensión para el navegador de Google Chrome?

Una vez hayan sido contestadas se dará paso a la prueba dividida en dos partes, una primera prueba para consultas simples y una segunda para consultas complejas, considerando frases nominales.

Las consultas definidas serán las mismas para todos los participantes con el fin de poder disponer de una base más sólida en cuanto a usabilidad e impresiones con la extensión. Además, como el propósito de este proyecto es otorgar al usuario de una mayor protección para sus datos más sensibles, las consultas seleccionadas serán

aquellas que impliquen búsquedas relacionadas con el estado de salud, situación económica o interés político.

Por otro lado, estas mismas consultas serán llevadas a cabo por uno de los dos tipos disponibles que ofrece la herramienta: privado y personalizado. En este caso, se les ofrece a los participantes manifestar que opción escogerían para realizar cada una de las consultas propuestas. La finalidad de esta parte de la prueba es realizar unas consultas más realistas y tener una perspectiva mejor de qué temas consideran los usuarios como sensibles, así en el caso de desarrollar una segunda versión se podría dar más énfasis en dichos aspectos, pudiendo hacer uso de ontologías específicas como MeSH especializada en el área médica.

Por último, se le pide al usuario que puntúe la extensión en términos de usabilidad, así como alguna anotación sobre la mejora de la extensión y sensaciones frente a los dos tipos de consultas disponibles.

5.4.1. Usuarios

La fase de pruebas es fundamental en cualquier proyecto de desarrollo, siendo muy útiles para evaluar la viabilidad de la herramienta. Estas pruebas serán realizadas con un total de cinco usuarios, ya que se considera un máximo de diez personas como óptimo para comprobar los diferentes problemas que puede proporcionar la herramienta creada.

Los participantes han sido escogidos bajo el criterio de obtener un conjunto de personas relativamente homogéneas, es decir, todos los participantes son adultos que trabajan en su día a día con ordenadores de escritorio. La razón de ello es que al no poder cuantificar de manera exacta la extensión, los participantes deberán dar su opinión respecto a la calidad de la respuesta resultante de la generalización, siendo esta muy subjetiva.

Para concluir con este apartado, se especifica el perfil de cada usuario dando a conocer el empleo que ejercen, pero identificados de forma numérica para no comprometer su privacidad.

- Usuario 1: Geógrafo.
- Usuario 2: Visual manager.
- Usuario 3: Community manager.
- Usuario 4: Enfermera.
- Usuario 5: Ingeniero superior de telecomunicaciones.

5.4.2. Resultados consultas simples

La prueba con consultas simples está formada por un total de 6 palabras que el usuario deberá buscar con la ayuda de la extensión y posteriormente, evaluar la calidad de la respuesta proporcionada. Esta evaluación se hace de forma cuantitativa, los participantes deben indicar un valor entre el 1 y el 5, siendo 1 el valor más bajo y, por lo tanto, no encuentra satisfactorias las respuestas encontradas o bien, un valor de 5 cuando la respuesta cumple las expectativas de búsqueda del usuario.

A continuación, se muestran dos tablas donde se observan las diferentes puntuaciones dadas por los participantes. En la primera tabla encontramos aquellas

consultas que los usuarios han considerado como privadas, mientras que la segunda tabla se ha realizado seleccionando el tipo de consulta personalizada.

- Consultas privadas

Consultas	**Usuario 1**	**Usuario 2**	**Usuario 3**	**Usuario 4**	**Usuario 5**
Herpes	5	5	x	x	x
Racism	2	x	3	3	x
Bets	x	5	x	5	5
Cocaine	5	4	5	5	4
Puntuación final	**4**	**4,6**	**4**	**4,3**	**4,5**

Tabla 3 Resultados consultas simples privadas

- Consultas personalizadas

Consultas	**Usuario 1**	**Usuario 2**	**Usuario 3**	**Usuario 4**	**Usuario 5**
Herpes	x	x	4	5	5
Anxiety	3	3	2	3	2
Racism	x	5	x	x	5
Dictatorship	5	5	3	4	4
Bets	5	x	5	x	x
Puntuación final	**4,3**	**4,3**	**3,5**	**4**	**4**

Tabla 4 Resultados consultas simples personalizadas

En las tablas se muestra una media relativa a las valoraciones aportadas por cada uno de los participantes. A raíz de estos resultados podemos observar cómo los participantes se encuentran satisfechos con las consultas realizadas, donde las consultas personalizadas han obtenido puntuaciones más bajas. Este hecho puede deberse a que la semántica utilizada es mediante sinónimos de uso menos común, proporcionando un resultado más alejado.

En cuanto al tipo de consulta, podemos observar qué consultas son percibidas como más sensibles y, por lo tanto, escogen realizarla de forma privada. En este caso, consultas sobre drogas o apuestas han tomado mayor relevancia que otras como la ansiedad o la dictadura.

Puesto que me resultaba curioso el encontrar que un dato relacionado con la salud mental lo considerasen como una consulta personalizada, pregunté a los participantes el porqué de su elección para tenerlo en cuenta si hacía falta en un futuro. Sus respuestas acerca de su elección con el tipo de consulta personalizada fueron qué preferían tener consultas concretas para disponer de información más exacta, que no el hecho de que se hiciera un perfil afín a su estado de salud mental.

5.4.3. Resultados consultas complejas

En este apartado se lleva a cabo la prueba que analiza la viabilidad de la extensión para consultas complejas. El procedimiento seguido es el mismo que en las consultas anteriores. Los participantes reciben una serie de consultas diferentes, las cuales deberán puntuar del 1 al 5 la calidad de la respuesta obtenida para cada uno del tipo de consulta que ellos crean conveniente.

En la primera tabla se tratan las consultas que los usuarios han considerado como privadas, mientras que en la segunda tabla encontramos aquellas que son personalizadas.

- Consultas privadas

Consultas	**Usuario 1**	**Usuario 2**	**Usuario 3**	**Usuario 4**	**Usuario 5**
Causes of AIDS	5	5	x	x	x
Heroin addiction	5	5	4	4	4
Anarchist party of America	x	x	5	5	x
Abortion hospital	5	5	5	5	5
Puntuación final	**5**	**5**	**4,6**	**4,6**	**4,5**

Tabla 5 Resultados consultas complejas privadas

- Consultas personalizadas

Consultas	**Usuario 1**	**Usuario 2**	**Usuario 3**	**Usuario 4**	**Usuario 5**
Causes of AIDS	x	x	5	5	5
Breast cancer symptoms	5	5	5	5	5
American marijuana laws	5	4	4	5	5
Anarchist party of America	1	3	x	x	2
Puntuación final	**3,6**	**4**	**4,6**	**5**	**4,25**

Tabla 6 Resultados consultas complejas personalizadas

Una vez calculada la puntuación media para cada uno de los usuarios, podemos observar cómo los participantes puntúan mejor las consultas complejas que las simples independientemente del tipo de consulta. Una posible razón es que al otorgar más palabras se realiza una consulta más específica con un tema y aunque se generalice cada una de ellas, el resultado obtenido sigue guardando una relación estrecha con la original.

Por otro lado, este mismo hecho ha supuesto que la mayoría de los participantes coincidan con el tipo de consultas a seleccionar. Si observamos los resultados,

podemos afirmar como las consultas que indican un estado de salud más comprometido como la adicción a la heroína o la búsqueda de un hospital para abortar, son consideradas como consultas de carácter sensible y, por lo tanto, se realizan de forma privada. En cambio, consultas orientadas a nivel político-social no representan un problema para los participantes.

5.4.4. Conclusiones pruebas con usuarios

Una vez se han analizado según el nivel de complejidad de las consultas, se extrae la media global tanto para consultas privadas como personalizadas por cada uno de los participantes.

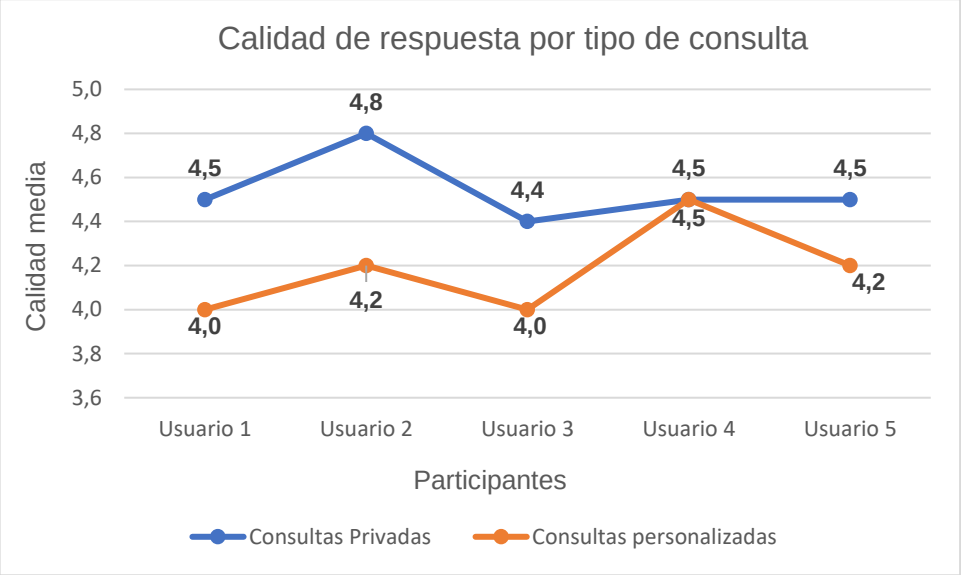


Ilustración 15 Gráfico calidad de respuesta por tipo de consulta

Como se muestra en el gráfico, las consultas privadas han sido muy satisfactorias con la generalización seleccionada, las puntuaciones de los usuarios han rondado entre el 4 y el 5. En cambio, la decisión semántica tomada para las consultas personalizadas han resultados ser un error, puesto que han proporcionado resultados con una relación más alejada de la esperada para este tipo de consulta. Aun así, ambos tipos de consultas han conseguido una puntuación por encima de la media, por lo que se puede afirmar haber cumplido el objetivo de mantener búsquedas que sean de interés para el usuario.

En relación con las preguntas mencionadas al principio de la sección, estas tenían como finalidad descubrir el nivel de conocimiento respecto a tecnologías de mejora de la privacidad, así como, si era necesario incluir un manual para hacer uso de la extensión. Con ellas se pudo conocer que todos los usuarios hacían uso del mismo tipo de herramientas obteniendo como resultado la pestaña de incógnito, la autenticación para la protección de datos y el uso de VPN en perfiles más técnicos, como es el caso del usuario 5.

Por otro lado, únicamente un participante sabía instalar directamente la extensión desde su ordenador, por ello a todos los participantes se les ofreció una guía que detalla los pasos a seguir para la instalación de la extensión en Google Chrome, una muestra de este manual se puede obtener en el apartado de Anexos del presente documento junto con las pruebas realizadas a cada uno de los participantes.

Respecto a las cuestiones efectuadas al terminar las pruebas, todos los participantes dieron una puntuación entre 4 y 5 sobre la facilidad de uso. De igual forma, todos coincidieron en utilizar la extensión únicamente para hacer consultas privadas que lleven datos más sensibles en ella.

En cuanto a la velocidad, se les pide a los usuarios que den sus impresiones a lo largo de la prueba, pero al tratarse de un valor muy subjetivo, no se puede llegar afirmar como de satisfactorias eran las respuestas. Mientras un usuario marcaba como negativo tener que cambiar a una segunda página para encontrar una respuesta acorde con la búsqueda, otros daban la misma situación como satisfactorio si al final obtenían su consulta. En cambio, sí que se pudo afirmar que el tiempo de respuesta para obtener la generalización fue muy positivo, notificando que no notaban grandes diferencias de tiempo entre hacer la consulta desde la extensión o directamente en Google.

Para concluir con el análisis, se les ofreció a los participantes la oportunidad de expresar aquellos aspectos que se podrían perfeccionar de cara a una segunda versión del producto en términos de interfaz y funcionalidad de la extensión. A continuación, se enumeran las propuestas aportadas:

- Mejorar la afinidad en algunas búsquedas privadas.
- Posibilidad de mantener la extensión fija en el navegador.
- Disponibilidad para más navegadores.
- Poder seleccionar qué palabras deseas que se generalicen.
- Disponibilidad de búsqueda en otro idioma.
- Opción de autocompletar en el buscador.

5.4.5. Rendimiento de la herramienta

Una vez se han realizado las correspondientes pruebas para determinar la calidad de las consultas junto con las primeras impresiones de los participantes, damos paso al último punto para determinar la viabilidad de la extensión.

Para ello, se va a hacer uso de una herramienta de Google Chrome, la cual permite a los usuarios conocer como están siendo distribuidos los recursos de dicho navegador. Esta herramienta se asemeja al administrador de tareas propio de Windows, un ejemplo de este gestor lo podéis encontrar en la siguiente página.

Este panel muestra qué procesos del navegador están consumiendo recursos. Dependiendo de si el usuario tiene o no pestañas abiertas, o existan extensiones instaladas en el navegador, se mostrarán en el administrador como tareas. Para cada una de ellas se indica la cantidad de memoria RAM que consumen, así como el uso de CPU, red e ID del proceso. En este caso, vamos a analizar únicamente los recursos relacionados con la memoria RAM y la CPU.

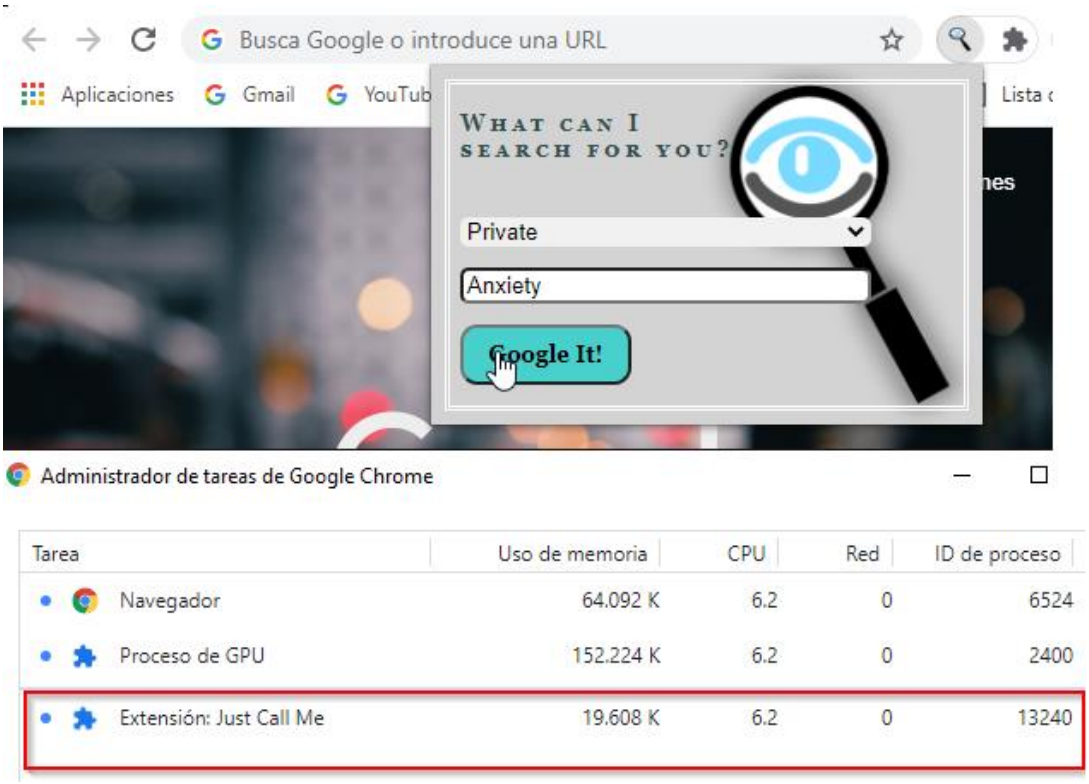


Ilustración 16 Gestor de tareas Google Chrome

Los datos mostrados en la figura son obtenidos en el momento que se está haciendo uso de la extensión, es decir, cuando un usuario indica el tipo y la consulta que desea realizar.

Como se puede comprobar, tanto en memoria como CPU el consumo es mínimo no llegando ni al 10% de la memoria. Se debe tener en cuenta, que en el caso de la CPU los valores difieren según la consulta, pero se mueven alrededor del 9 y 6%, pudiendo variar según el procesador que incorpore cada computador.

En cuestión de velocidad, se consigue un tiempo de respuesta óptimo cercano al de Google, este último ronda los 300ms en tiempo de respuesta [3]. No obstante, si el servidor se encuentra apagado, en el momento que se vuelva a conectar, la primera consulta se demorara más hasta realizar la conexión. Una vez la conexión se haya establecido de forma correcta, el resto de las consultas se harán rondando los 700ms aproximadamente, siendo un tiempo de respuesta válido teniendo en cuenta que es necesario transformar la consulta del usuario en una nueva a través de la función en Python.

5.5. Futuro trabajo

A lo largo del proyecto se ha podido comprobar la carencia de herramientas necesarias para mejorar la privacidad del usuario.

Nuestro prototipo actual se encuentra implementado en el lado del cliente, otorgándole al usuario una mayor protección, ya que permite realizar búsquedas más genéricas y como consecuencia, se crea un perfil de intereses con datos menos sensibles. Sin embargo, la tecnología se encuentra en continuo desarrollo y por ello es conveniente realizar futuras mejoras para su funcionalidad.

A raíz de las aportaciones de los usuarios mencionadas anteriormente junto con el análisis de datos realizado, se puede optar por dos vías: o bien seguir con la extensión y hacer uso de la herramienta como hasta ahora o, por otro lado, pasar la funcionalidad a una aplicación para dispositivos móviles.

Si se decide seguir con la extensión, es conveniente mejorar la conexión al servidor. Actualmente se encuentra implementada mediante un servidor simple, puesto que el objetivo principal era demostrar la viabilidad de la herramienta. No obstante, lo primero que se debería tomar en cuenta en una segunda versión es complementar el servidor Gunicorn con un segundo servidor que otorgue mayor seguridad. Una posibilidad sería Nginx, ya que proporciona la funcionalidad de trabajar como un proxy inverso de cliente y ofrecer mayor protección SSL[14] en las consultas.

Una segunda mejora respecto a la extensión es poder realizar la consulta directamente desde el buscador de Google. En esta primera versión, la extensión abre una nueva pestaña cada vez que se realiza una búsqueda, siendo bastante molesto cuando el usuario desea realizar más de una consulta seguida.

El supuesto de pasar la extensión a una aplicación móvil resulta más interesante, ya que la mayoría de los usuarios de Internet realizan las consultas mediante estos dispositivos. Esta aplicación incluiría las mismas funcionalidades que la extensión, con la ventaja que el usuario tendría más facilidades al ser ejecutado todo desde su teléfono personal.

En cuanto a la funcionalidad de la herramienta, se mejoraría la búsqueda para las consultas personalizadas cambiando el criterio escogido sobre la transformación de la consulta. Por otro lado, en el caso de las consultas complejas se podría añadir nuevas funcionalidades como la selección de palabras que quieran generalizar, así como incorporar frases que contengan verbos.

Por último, se debería estudiar incluir WordNet en español, así como utilizar otros directorios que aporten más palabras con el fin de obtener una base de datos más extensa y mejorar la afinidad de las búsquedas.

---

[14] Secure Sockets Layer

6. Conclusiones

El campo de la seguridad informática se encuentra en pleno auge dentro de nuestra sociedad, cada vez son más las personas que se conciencia sobre su huella digital y la importancia de no administrar sus datos a cualquier entidad. Al principio de este documento se hablaba de los diferentes servicios que se ofrecen en la web, uno de los más antiguos y que no para de crecer son las consultas realizadas en motores de búsqueda como es el caso de Google el buscador utilizado en este proyecto.

A pesar de que este buscador mantiene cierta transparencia respecto a la recogida de datos y el seguimiento de cookies, no especifica qué hace realmente con esos datos ni como los tiene almacenados, provocando que se cree un perfil de intereses con todo tipo de datos personales, incluidos aquellos que son considerados sensibles. Este hecho motivó a investigar este campo, desde que mecanismos y tecnologías existen, hasta desarrollar una herramienta que permite crear perfiles más genéricos en cuanto a intereses.

En la primera parte del proyecto se llevó a cabo un estudio sobre qué mecanismos y tecnologías existían para proteger o mejorar la privacidad de los datos. En el caso de los mecanismos se pudo comprobar como el objetivo principal de muchos autores es buscar el anonimato con la ayuda de una entidad externa. El problema de esta clase de mecanismos es que requieren confiar plenamente en esta entidad, quien podría acabar utilizando los datos sensibles de forma malintencionada. En cambio, los mecanismos de perturbación que proporcionan un mayor control al usuario apenas son llevados a la práctica, ya sea por falta de recursos o bien por falta de interés en esta área. Por otro lado, en el estudio de mercado sobre las tecnologías de mejora, si bien en la actualidad existen varias tecnologías como TrackMeNot y GooPIR que permiten enviar consultas falsas a los buscadores, estas no evitaban que se perfilara a los usuarios.

Una vez se dio por finalizada la parte de investigación se llegó a la conclusión que no existían herramientas suficientes para hacer frente a la recolecta de datos que realizan los buscadores, en este proyecto en concreto, únicamente se encontró las dos tecnologías mencionadas anteriormente. Este hecho motivó a que el mecanismo implementado en este proyecto fuese la generalización de consultas y así, comprobar qué viabilidad de futuro tendría una herramienta de este tipo para otorgar al usuario una mayor protección de sus datos.

A partir de lo expuesto, se realizó una segunda fase donde dio lugar al desarrollo de la herramienta marcando como objetivo la creación de una tecnología que permitiera realizar consultas más genéricas a partir de consultas de carácter sensible.

La herramienta desarrollada ha resultado ser una extensión para el navegador de Chrome siendo totalmente funcional y disponible para el motor de búsqueda Google, capaz de realizar tanto consultas simples como complejas, además de devolver una consulta totalmente nueva pero semánticamente relacionada a la original gracias a WordNet, la base de conocimiento seleccionada.

A pesar de ser una primera versión con unas funcionalidades básicas para cumplir los objetivos marcados al inicio de este proyecto, se ha podido observar que se trata de una tecnología que carece de ciertas necesidades que el usuario ve esencial para poder hacer uso de ella, como puede ser una mayor afinidad entre la consulta original

y la consulta nueva. Por este motivo, es conveniente mejorar la relación semántica para cada uno de los dos tipos de consultas creados, dando más énfasis en las personalizadas o bien, tomando otro camino diferente como puede ser otorgar únicamente consultas privadas que se especialicen más en la generalización.

En un primer momento se había planteado la posibilidad de crear una aplicación para dispositivos móviles que ofreciera a los usuarios realizar consultas a partir de él. Sin embargo, un imprevisto en la planificación supuso tener que cambiar la herramienta planteada por una extensión. El principal problema en la planificación vino dado por la falta de concordancia que mantienen los autores de este campo respecto algunos términos empleados, como es el caso de mecanismo y tecnología. Esta confusión supuso tener que realizar una segunda investigación respecto a los dos primeros apartados que provocó un atraso en la planificación fijada al inicio del proyecto.

A pesar de los contratiempos encontrados, en la segunda fase de este proyecto se consiguió alcanzar la planificación marcada. Si es cierto que hubo dificultades en el momento de desarrollar la conexión entre servidor y usuario, al tratar con una metodología ágil, permitió ir avanzando en cada fase a medida que se tomaban decisiones de implementación. Este proceso se llevaba a cabo cada vez que se daban nuevos conflictos, lo que suponía ir modificando o encontrando nuevos requisitos o funciones según iba demandando la herramienta. Además, al tratar con este tipo de metodología permitía realizar las pruebas con usuarios una vez se encontraba operativa, a pesar de poder estar cambiando aspectos relacionados con el diseño gráfico de forma paralela.

Por otra parte, han quedado aspectos que hubieran sido interesantes de tratar en el desarrollo de la herramienta como, por ejemplo, permitir que la extensión funcione desde el navegador de Google. En un primer momento, se consideró diferenciar la extensión del propio buscador con la ayuda de una ventana emergente que podía ser utilizada desde cualquier ventana del navegador. Pero a raíz de las pruebas con usuarios, se comprobó que resultaba molesto tener que ir abriéndola en el caso de estar haciendo más de una consulta seguida.

Otro aspecto interesante que se habría querido tratar es la función de eliminar el historial de búsqueda una vez el usuario ha terminado de utilizar el navegador. Esta característica se podría llevar a cabo gracias a la API de Google que incluye un método para ello, desgraciadamente los tiempos no permitieron investigar esta línea de trabajo y únicamente se implementó los objetivos principales que se habían determinado al inicio del proyecto.

Para finalizar, pongo de manifiesto una pequeña crítica surgida a raíz de ir haciendo este proyecto. Si es cierto que los usuarios de internet son conscientes de las vulnerabilidades a las que están expuestos, ya sean virus, falsificación de identidades o el seguimiento de las cookies, no dan la importancia que requiere el dar sus datos privados a una entidad como Google. Esto conlleva una falta de concienciación social que impide desarrollar tecnologías más estrictas en relación con los datos compartidos, ya que el principal problema que un usuario le exigirá es tener lo que busca en poco tiempo y de forma correcta, y este requisito tiene un precio. Por ello, si se carece de la disponibilidad de recursos para poder llevar esto a cabo, lo mínimo es ejercer una presión por parte de la ciudadanía a las grandes empresas de la red para que se cree una mayor transparencia en cómo tratan nuestros datos.

7. Glosario

anonimato *9*, *10*, *12*, *13*, *14*, *38*

Propiedad que define la acción de ocultar el vínculo entre una identidad y una acción o información.

API *22*, *39*

Conjunto de especificaciones de comunicación entre componentes software para intercambiar datos.

backend *28*

Programa que procesa la información de una página web siendo oculto para el usuario como, por ejemplo, la comunicación con el servidor.

base de conocimiento *3*, *4*, *6*, *17*, *21*, *38*

Base de datos centralizada que permite recopilar, organizar, buscar y compartir información y datos.

cifrado de claves *9*

Transformación de un texto en claro, mediante un algoritmo que tiene como parámetro una clave, en un texto cifrado ininteligible para quien no conozca la clave del descifrado.

cifrado homomórfico *9*

Esquema de cifrado que permite la posibilidad de trabajar con datos cifrados sin necesidad de descifrarlos, minimizando la posibilidad de exposición de la información.

clave privada *9*

Clave que permite realizar la transformación criptográfica inversa a la que se obtiene con una clave pública siendo computacionalmente inviable obtener a partir de esta última.

clave pública *9*, *14*

Clave que permite realizar la transformación criptográfica inversa a la que se obtiene con una clave privada y que se puede obtener fácilmente a partir de esta última.

comunicación P2P *10*

Comunicación de igual a igual. Modelo de comunicación descentralizado donde cada parte o usuario actúan por igual y pueden tener la función de servidor o de cliente.

contenedor *29*

Tecnologías que le permiten empaquetar y aislar las aplicaciones junto con todos los archivos que requieren para ejecutarse. Esto permite mover la aplicación que se encuentra dentro del contenedor entre los diferentes entornos sin perder ninguna de sus funciones.

cookies *2*

Archivo de texto pequeño que se almacena en el disco duro y que permite que un sitio Web haga un seguimiento de la asociación del usuario a dicho sitio.

cortafuegos *16*

Sistema de seguridad para bloquear accesos no autorizados a un ordenador mientras sigue permitiendo la comunicación con otros servicios autorizados.

criptografía *9*, *14*

Técnica o conjunto de métodos cuya función es transformar un determinado mensaje o información en otro totalmente distinto e ininteligible por cualquier persona que no esté autorizada a leerlo.

Delicious *11*

Servicio de gestión de marcadores sociales en web.

enrutamiento de cebolla *14*

Método mediante el cual los paquetes de red se pueden transmitir de forma anónima a través de Internet.

etiquetado colaborativo *8*, *11*

Plataforma de comunicación en línea que permite almacenar, clasificar y compartir enlaces en Internet pudiendo existir servicios especializados en diferentes áreas como libros, música o compras.

extensión *14*, *21*,*46*,*50*

Pequeños programas que se instalan dentro del navegador y añaden o mejoran algunas de sus funciones.

falsificación de datos *11*

Método donde se proporciona información que no reflejan los datos originales.

frontend *22*

Parte de una web que conecta e interactúa con los usuarios que la visitan siendo la parte correspondiente al diseño, los contenidos y la que permite a los visitantes navegar por las diferentes páginas.

generalización de consultas *3*, *4*, *6*, *12*, *16*, *20*, *38*

Mecanismo que reemplaza un conjunto de datos originales en un conjunto de datos más genéricos.

Git *29*

Herramienta que permite hacer modificaciones en el código y hace que sea más fácil la administración de las distintas versiones de cada producto desarrollado.

hiperónimo *17*, *27*

Palabra cuyo significado está incluido en el de otras.

HTTP *15*, *25*, *28*

Protocolo de comunicación que permite las transferencias de información a través de archivos en la World Wide Web.

huella digital *16*, *38*

Rastro de datos que se crea mientras un usuario navega por internet, incluyendo sitios web que visita y la información que envía a los servicios en línea.

lenguaje natural *17*

Lenguaje utilizado por los seres humanos para comunicarse entre ellos.

mecanismos de protección de datos *7*

Proceso, técnica o sistema para conseguir un mayor control de los datos.

MeSH *31*

Diccionario de sinónimos de vocabulario controlado por NLM que se utiliza para indexar artículos para PubMed.

método GET *25*

Método HTTP que envía la información codificada del usuario en la cabecera de petición HTTP, directamente en la URL.

metodología ágil *3*

Tipo de enfoque para la gestión de proyectos donde se tiene en cuenta las actividades por hacer, trabajando de forma incremental e iterativa.

ODP *3*

Directorio web para listar y categorizar enlaces a páginas web.

ontologías *8*

Esquema de representación que especifica de forma explícita y compartida la información de un dominio.

perfiles de usuarios *2*

Recopilación y análisis de los datos que los usuarios comunican a los sistemas de recomendación o motores de búsqueda.

privacidad *hard* *13*, *15*

Permite a los usuarios proteger su privacidad frente a diferentes servicios sin la necesidad de confiar en entidades externas.

privacidad *soft* *13*

La protección de los datos viene dada por una entidad externa, no habrá control de la información por parte del usuario.

proxy *13*, *14*, *37*

Servidor que se encarga de realizar las conexiones solicitadas con el exterior i retransmitirlas hacia el equipo que había iniciado la conexión.

servicios basados en la localización *10*

Servicio de información que utiliza datos de ubicación del usuario para controlar sus prestaciones y características, gracias a tecnologías relacionadas con la información geográfica, tales como sistemas GIS, GPS, o Wi-Fi.

servidor HTTP *28*

Programa informático que procesa una aplicación del lado del servidor, realizando conexiones bidireccionales o unidireccionales y síncronas o asíncronas con el cliente y generando o cediendo una respuesta en cualquier lenguaje o aplicación del lado del cliente.

sistemas de recomendaciones *8*

Herramienta que establece un conjunto de criterios y valoraciones sobre los datos de los usuarios para realizar predicciones sobre recomendaciones de elementos que puedan ser de utilidad o valor para el usuario.

SSL *37*

Protocolo complementario al protocolo de Internet que tiene como objetivo retransmitir la información de manera segura mediante el cifrado de la misma.

supresión de datos *11*

Técnica que permite abstenerse de proporcionar cierta información con el fin de proporcionar datos menos precisos.

tecnologías de comunicación anónimas *13*

Tecnología que oculta la dirección IP de un usuario del servidor que aloja el sitio web visitado por el usuario.

tecnologías de mejora de la privacidad *7*

Tecnologías a nivel de usuario diseñadas con el fin de cumplir su objetivo sin poner en riesgo la privacidad y seguridad de las personas que hacen uso de él.

tesauro *15*

Lista de palabras o términos controlados, empleados para representar conceptos.

VPN *14*, *34*, *50*

Herramienta mediante la cual es posible hacer que un ordenador remoto se comporte como si estuviera ubicado físicamente dentro de la organización.

Web Semántica *8*, *11*

Web extendida, dotada de mayor significado en la que cualquier usuario en Internet, podrá encontrar respuestas a sus preguntas de forma más rápida y sencilla gracias a una información mejor definida.

Wireframe *22*

Esquema de página que se utiliza como guía visual para representar la estructura esquelética de un sitio web.

WordNet *3*, *6*, *8*, *17*, *18*, *21*, *23*, *36*

Base de datos léxica de inglés formada por sustantivos, verbos, adjetivos y adverbios agrupados en conjuntos de sinónimos denominados *synsets* que expresan un concepto específico.

WSGI *28*

Especificación que describe como los servidores web reenvían las solicitudes a aplicaciones web o marcos escritos en el lenguaje de programación Python.

8. Bibliografía

[1] S. Kemp, «Digital 2021: Global Overview Report,» Date Reportal, 2021. [En línea]. Available: https://datareportal.com/reports/digital-2021-global-overview-report. [Último acceso: 18 03 2021].

[2] D. S. Alexandre Viejo, «Profiling Social Networks to Provide Useful and Privacy-Preserving Web Search,» *Journal of the Association for Information Science and Technology,* vol. 65, nº 12, pp. 2444-2458, 2014.

[3] J. C.-R. A. V. David Sánchez, «Knowledge-Based Scheme to Create Privacy-Preserving but Semantically-Related Queries for Web Search Engines,» *Information Sciences,* vol. 218, pp. 17-30, 2013.

[4] Y. &. K. A. Wang, «Privacy-enhancing technologies,» de *Handbook of Research on Social and Organizational Liabilities in Information Security*, Hershey, PA, IGI Global, 2008, pp. 203-227.

[5] J. Borkin, Why Adopting Privacy Enhancing Technologies (PETs) Takes so Much Time, S. Gutwirth, Y. Poullet, P. Hert, R. Leenes, 2011, pp. 309-341.

[6] E. U. A. f. N. a. I. Security, «Readiness Analysis for the Adoption and Evolution of Pri-vacy Enhancing Technologies,» Science and Technology Park of Crete (ITE), Heraklion, 2015.

[7] J. H. O. L. T. Berners-Lee, «The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities,» *Scientific American,* vol. 284, nº 5, pp. 34-43, 2001.

[8] L. Sweeney, «k-anonymity: A model for protecting privacy,» *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* vol. 10, nº 5, p. 557–570, 2002.

[9] D. Chaum, «Untraceable electronic mail, return addresses, and digital,» *Communications ACM,* vol. 24, nº 2, p. 84–90, 1981.

[10] W. Ahmad y A. Khokhar, «An architecture for privacy preserving collaborative filtering on Web portals,» *Proceedings of the IEEE International Symposium on Information Assurance and Security,* p. 273–278, 29-31 Agosto 2007.

[11] C. Chow, M. Mokbel y X. Liu, «A peer-to-peer spatial cloaking algorithm for anonymous location-based services,» *Proceedings of the ACM International Symposium Advances Geographic Information Systems (GIS), Arlington,* p. 171–178., 10–11 Noviembre 2006.

[12] A. V. J. H.-J. J. Castellà-Roca, «Preserving user's privacy in web search engines,» *Computer Communications,* vol. 32, nº 13–14, p. 1541–1551, 2009.

[13] D. R.-M. J. F. J. Parra-Arnau, «A privacy-preserving architecture for the semantic Web based on tag suppression.,» *Proceedings International Conference Trust, Privacy, Security, Digital Business (TrustBus),* vol. 6264, nº 58–68, p. 58–68, 30 –3 Agosto-Septiembre 2010.

[14] A. P. E. F. J. F. D. R.-M. J. Parra-Arnau, «Privacy-Preserving Enhanced Collaborative Tagging,» *IEEE Trans. Knowl. Data Eng.,* vol. 26, nº 1, pp. 180-193, 2014.

[15] D. R.-M. J. F. J. Parra-Arnau, «Optimal Forgery and Suppression of Ratings for Privacy Enhancement in Recommendation Systems,» *(MDPI) Entropy,* vol. 16, nº 3, pp. 1586-1634, 2014.

[16] P. Samarati, «Protecting Respondents' Identities in Microdata Release,» *IEEE Transactions on Knowledge and Data Engineering,* vol. 13, nº 6, pp. 1010-1027, 2001.

[17] M. DENG, «Privacy Preserving Content Protection,» Ph.D. Dissertation, Katholieke University, Lovaina,Bélgica, Junio 2010.

[18] «Kproxy,» [En línea]. Available: https://kproxy.com/. [Último acceso: 03 05 2021].

[19] P. F. S. a. D. M. G. Michael G. Reed, «Proxies for anonymous routing.,» de *Proceedings of the Computing Security Application Conference (CSAC)*, San Diego, CA, 1996, p. 9–13.

[20] «Tor Project,» [En línea]. Available: https://www.torproject.org/. [Último acceso: 03 05 2021].

[21] «I2P,» [En línea]. Available: https://geti2p.net/en/. [Último acceso: 30 05 2021].

[22] «Private tunnel,» [En línea]. Available: https://www.privatetunnel.com/. [Último acceso: 30 05 2021].

[23] «NordVPN,» [En línea]. Available: https://nordvpn.com/es/. [Último acceso: 30 05 2021].

[24] «TrackMeNot,» [En línea]. Available: https://trackmenot.io/. [Último acceso: 03 05 2021].

[25] H. N. D. C Howe, «TrackMeNot: Resisting Surveillance in Web Search,» de *On the Identity Trail: Privacy, Anonymity and Identity in a Networked Society, Eds.*, Oxford, Oxford University Press, 2009, p. 23.

[26] «GooPIR,» [En línea]. Available: https://www.portalprogramas.com/goopir/.

[27] A. S. J. C.-R. J. Domingo-Ferrer, «h(k)-private information retrieval from Privacy-Uncooperative,» *Journal of Online Information Review,* vol. 33, nº 4, pp. 1468-4527, 2009.

[28] N. S. S. T. Peddinti, «On the privacy of web search based on query obfuscation: a case study of TrackMeNot,» de *Proceedings of the 10th international conference on Privacy enhancing technologies – PETS'10*, 2010, pp. 19-37.

[29] «Startpage,» [En línea]. Available: https://www.startpage.com/. [Último acceso: 30 05 2021].

[30] «Gibiru,» [En línea]. Available: https://gibiru.com/. [Último acceso: 30 05 2021].

[31] «WordNet,» [En línea]. Available: https://wordnet.princeton.edu/. [Último acceso: 05 04 2021].

[32] P. University, About WordNet, WordNet Princeton University, 2010.

[33] «Chrome Developers,» [En línea]. Available: https://developer.chrome.com/docs/extensions/. [Último acceso: 30 05 2021].

[34] «Heroku,» [En línea]. Available: https://www.heroku.com/. [Último acceso: 30 05 2021].

[35] «Data Protection and Privacy,» 2021. [En línea]. Available: https://cloudian.com/guides/data-protection/data-protection-and-privacy-7-ways-to-protect-user-data/#protection-technologies.

[36] D. R.-M. J. F. Javier Parra-Arnau, «Measuring the Privacy of User Profiles in Personalized Information Systems,,» vol. 33, Elsevier Future Gen. Comput. Syst. (FGCS), Special Issue Data Knowl. Eng., 2014, pp. 53-63.

[37] J. Sartain, «Top 5 tools to protect internet privacy,» 2017. [En línea]. Available: https://www.csoonline.com/article/3213931/top-5-tools-to-protect-internet-privacy.html. [Último acceso: 03 05 2021].

[38] DrSoft, «The Most Private Search Engines,» 13 11 2020. [En línea]. Available: https://drsoft.com/2019/11/01/the-most-private-search-engines/. [Último acceso: 30 05 2021].

[39] Google, «Privacidad y condiciones,» [En línea]. Available: https://policies.google.com/. [Último acceso: 30 05 2021].

9. Anexos

9.1. Manual de instalación de la extensión

Para poder realizar correctamente la instalación de la extensión, es necesario tener guardada la carpeta con el código de implementación en su dispositivo. Una vez tenga localizada la carpeta, el procedimiento a seguir es muy sencillo, constando únicamente de 4 simples pasos:

1. Dirígete a **chrome://extensions**. Si su navegador ya dispone de una extensión podrá ir directamente al dar clic sobre el puzle que encontrará en el margen superior derecho y seleccionando la opción gestionar extensiones.

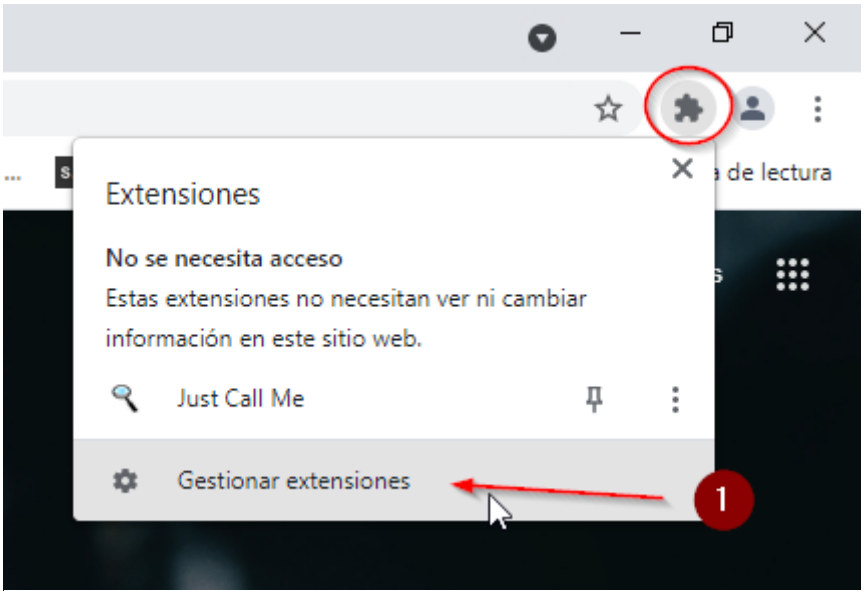


Ilustración 17 Primer paso: chrome://extensions

2. Dentro de las extensiones, activa el **modo de desarrollador** situado en la parte superior derecha.

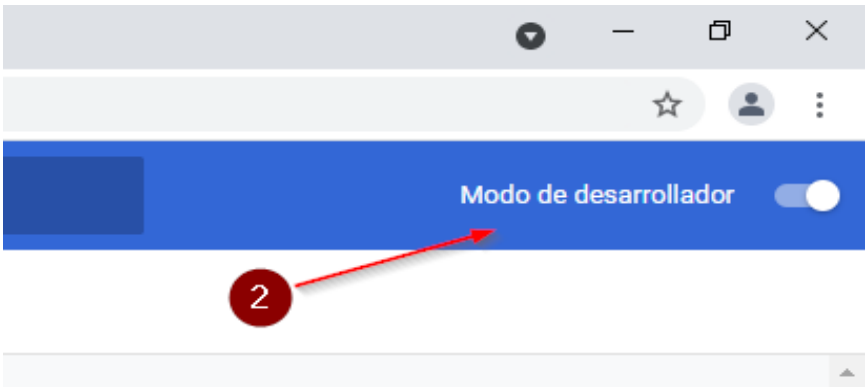


Ilustración 18 Segundo paso: modo desarrollador

3. Al activar el botón anterior aparecerá la opción de cargar un paquete descomprimido, **cargar descomprimida**. Haz clic en él.

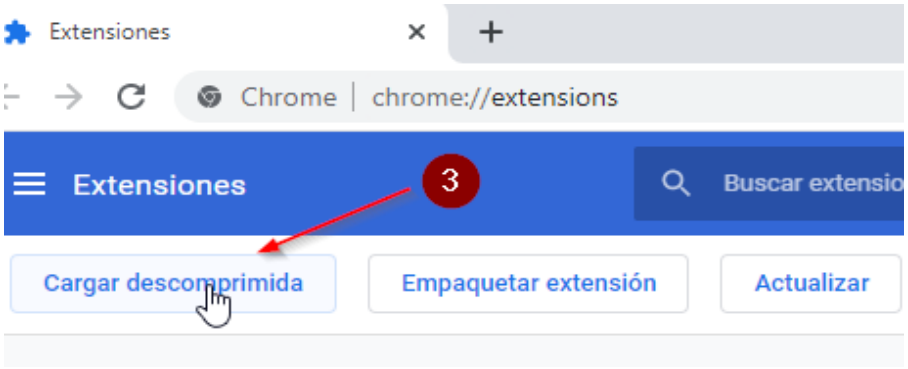


Ilustración 19 Tercer paso: cargar descomprimida

4. Busca y selecciona la carpeta dentro de tu dispositivo. Una vez cargada, la extensión se mostrará en el catálogo de tus extensiones.

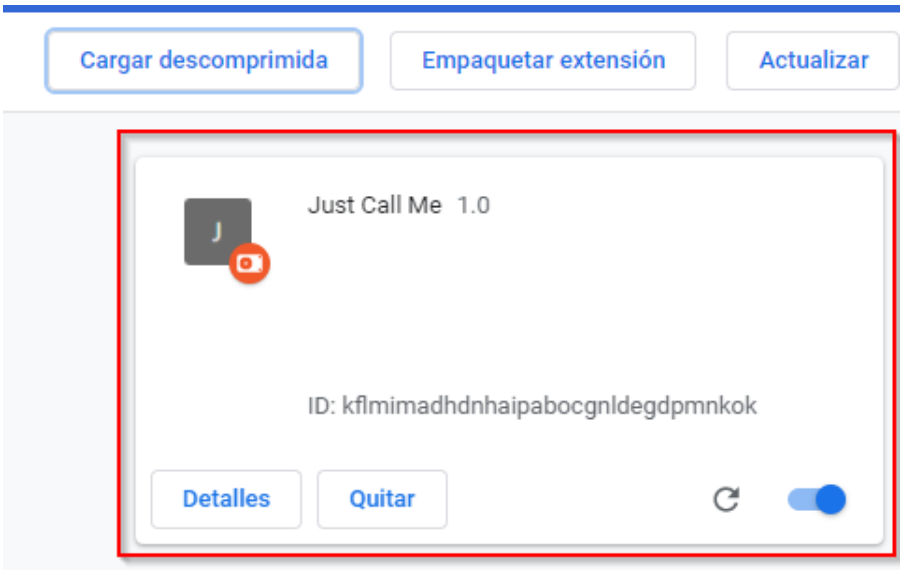


Ilustración 20 Extensión instalada

9.2. Pruebas con usuarios

9.2.1. Usuario nº1

- **¿Toma alguna medida para proteger sus datos privados? En caso afirmativo, podría decirme cuál.** Autenticación y pestaña de incógnito.
- **¿Sabría instalar una extensión para el navegador de Google Chrome?** Si.

Consultas	Tipo de consulta	Calidad de la consulta
Herpes	Privada	5
Anxiety	Personalizada	3
Racism	Privada	2
Dictatorship	Personalizada	5
Bets	Personalizada	5
Cocaine	Privada	5
Causes of AIDS	Privada	5
Heroin problems	Privada	5
Breast cancer symptoms	Personalizada	5
American marijuana laws	Personalizada	5
Anarchist party of America	Personalizada	1
Abortion hospital	Privada	5

Tabla 7 Resultados usuario nº1

- Facilidad de uso de la extensión: 4.
- Tipo de consulta: Privadas.
- Aspectos que mejorar de la extensión: Mejorar afinidad en las búsquedas y reducir número de pestañas.

9.2.2. Usuario nº2

- **¿Toma alguna medida para proteger sus datos privados? En caso afirmativo, podría decirme cuál.** Pestaña de incógnito.
- **¿Sabría instalar una extensión para el navegador de Google Chrome?** No

Consultas	Tipo de consulta	Calidad de la consulta
Herpes	Privada	5
Anxiety	Personalizada	3
Racism	Personalizada	5
Dictatorship	Personalizada	5
Bets	Privada	5
Cocaine	Privada	4
Causes of AIDS	Privada	5
Heroin problems	Privada	5
Breast cancer symptoms	Personalizada	5
American marijuana laws	Personalizada	4
Anarchist party of America	Personalizada	3
Abortion hospital	Privada	5

Tabla 8 Resultados usuario nº2

- Facilidad de uso de la extensión: 4.
- Tipo de consulta: Privadas.
- Aspectos que mejorar de la extensión: Mantener fija la extensión en la parte derecha de la pantalla.

9.2.3. Usuario nº3

- **¿Toma alguna medida para proteger sus datos privados? En caso afirmativo, podría decirme cuál.** Autenticación y pestaña de incógnito.
- **¿Sabría instalar una extensión para el navegador de Google Chrome?** No

Consultas	Tipo de consulta	Calidad de la consulta
Herpes	Personalizada	4
Anxiety	Personalizada	2
Racism	Privada	3
Dictatorship	Personalizada	3
Bets	Personalizada	5
Cocaine	Privada	5
Causes of AIDS	Personalizada	5
Heroin problems	Privada	4
Breast cancer symptoms	Personalizada	5
American marijuana laws	Personalizada	4
Anarchist party of America	Privada	5
Abortion hospital	Privada	5

Tabla 9 Resultados usuario nº3

- Facilidad de uso de la extensión: 5.
- Tipo de consulta: Privadas.
- Aspectos que mejorar de la extensión: Elegir que palabras se generalizan, disponibilidad para más navegadores y futura aplicación.

9.2.4. Usuario nº4

- **¿Toma alguna medida para proteger sus datos privados? En caso afirmativo, podría decirme cuál.** Autenticación.
- **¿Sabría instalar una extensión para el navegador de Google Chrome?** No

Consultas	Tipo de consulta	Calidad de la consulta
Herpes	Personalizada	5
Anxiety	Personalizada	3
Racism	Privada	3
Dictatorship	Personalizada	4
Bets	Privada	5
Cocaine	Privada	5
Causes of AIDS	Personalizada	5
Heroin addiction	Privada	4
Breast cancer symptoms	Personalizada	5
American marijuana laws	Personalizada	5

Anarchist party of America	Privada	5
Abortion hospital	Privada	5

Tabla 10 Resultados usuario nº4

- Facilidad de uso de la extensión: 5.
- Tipo de consulta: Privadas.
- Aspectos que mejorar de la extensión: Permitir consultas en otros idiomas.

9.2.5. Usuario nº5

- **¿Toma alguna medida para proteger sus datos privados? En caso afirmativo, podría decirme cuál.** VPN, autenticación y pestaña de incógnito.
- **¿Sabría instalar una extensión para el navegador de Google Chrome?** No

Consultas	Tipo de consulta	Calidad de la consulta
Herpes	Personalizada	5
Anxiety	Personalizada	2
Racism	Personalizada	5
Dictatorship	Personalizada	4
Bets	Privada	5
Cocaine	Privada	4
Causes of AIDS	Personalizada	5
Heroin problems	Privada	4
Breast cancer symptoms	Personalizada	5
American marijuana laws	Personalizada	5
Anarchist party of America	Personalizada	2
Abortion hospital	Privada	5

Tabla 11 Resultados usuario nº5

- Facilidad de uso de la extensión: 5.
- Tipo de consulta: Privadas.
- Aspectos que mejorar de la extensión: Autocompletar la búsqueda.