ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA
GRADO EN INGENIERIA INFORMÁTICA

Diseño e Implementación de un robot de conocimiento para la extracción de información de los Social Media y Web

Design and implement a Knowbots for extraction of Social Media and Web's information.

Realizado por
Francisco Javier Gutiérrez Torres
Tutorizado por
Dr. José Ignacio Peláez Sánchez
Departamento
Lenguaje y Ciencias de la Programación

UNIVERSIDAD DE MÁLAGA
MÁLAGA, Septiembre 2015

Fecha defensa:
El Secretario del Tribunal

Resumen

Las transformaciones tecnológicas y de información que está experimentando la sociedad, especialmente en la última década, está produciendo un crecimiento exponencial de los datos en todos los ámbitos de la sociedad. Los datos que se generan en los diferentes ámbitos se corresponden con elementos primarios de información que por sí solos son irrelevantes como apoyo a las tomas de decisiones. Para que estos datos puedan ser de utilidad en cualquier proceso de decisión, es preciso que se conviertan en información, es decir, en un conjunto de datos procesados con un significado, para ayudar a crear conocimiento. Estos procesos de transformación de datos en información se componen de diferentes fases como la localización de las fuentes de información, captura, análisis y medición.

Este cambio tecnológico y a su vez de la sociedad ha provocado un aumento de las fuentes de información, de manera que cualquier persona, empresas u organización, puede generar información que puede ser relevante para el negocio de las empresas o gobiernos. Localizar estas fuentes, identificar información relevante en la fuente y almacenar la información que generan, la cual puede tener diferentes formatos, es el primer paso de todo el proceso anteriormente descrito, el cual tiene que ser ejecutado de manera correcta ya que el resto de fases dependen de las fuentes y datos recolectados.

Para la identificación de información relevante en las fuentes se han creado lo que se denomina, robot de búsqueda, los cuales examinan de manera automática una fuente de información, localizando y recolectando datos que puedan ser de interés.

En este trabajo se diseña e implementa un robot de conocimiento junto con los sistemas de captura de información online para fuentes hipertextuales y redes sociales.

Palabras Claves

Big Data, diario, Eclipse, Hibernate, SQL Server 2012, redes sociales, fuentes de información, interfaz, RSS, empresa, público, patrones, Java, php, javascript, Wamp Server, Notepad, Web 2.0.

Abstract

The transformations technology and information that is experiencing society, especially in the last decade, is producing an exponential growth of data in all areas of society. Data generated in different areas correspond with primary information that by themselves are irrelevant as support to decision-makers. So these data can be useful in any decision-making process, to become information, i.e., a set of processed data with a meaning, to help create knowledge needed. These processes of transformation of data into information consist of different phases as the location of the sources of information, capture, analysis and measurement.

This technological and at the same time of the society change has led to an increase of sources of information, so that any person, business or organization, can generate information that may be relevant to the business of companies or Governments. Locating these sources, identify relevant information at the source and store the information they generate, which can have different formats, it is the first step of the process referred to above, which have to be executed properly since the rest of phases depend on sources and collected data.

For the identification of relevant information in the sources they have created what is called, robot's search, which examines a source of information, automatically locating and collecting information that may be of interest.

In this work it is designed and deployed a robot of knowledge along with the systems for capturing information online for hipertext sources and social networks.

Keywords:

Big Data, journal, Eclipse, Hibernate, SQL Server 2012, social media, information, interface, RSS, company, public, patterns, Java, php, javascript, Wamp Server, Notepad, Web 2.0.

Índice del contenido

1. Introducción

1.1. Motivación del proyecto

Las transformaciones tecnológicas y de información que está experimentando la sociedad, especialmente en la última década, está produciendo un crecimiento exponencial de los datos en todos los ámbitos de la sociedad. Así por ejemplo, en el año 2012 el volumen de información se estimaba en terabyte, mientras que en el año 2015 el volumen de información se está estimando en zettabyte.

Este crecimiento de información ha dado lugar a lo que se denomina Big Data. Pero ¿Qué es Big Data? El Big Data (o Datos masivos) es un concepto que hace referencia a la acumulación masiva de datos y a los procedimientos usados para identificar patrones recurrentes dentro de esos datos. Otras denominaciones para el mismo concepto son datos masivos o datos a gran escala. En la figura 1 se muestra como los documentos que contienen el término 'Big Data' ha ido creciendo a los largo de los últimos años, y este crecimiento continua aumentando cada año más.

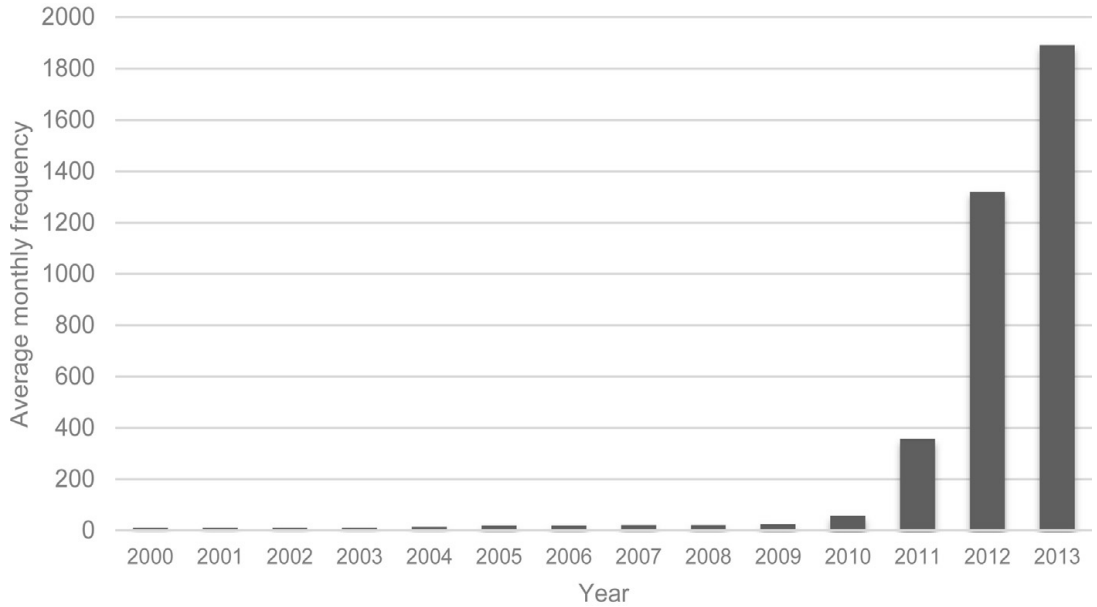

Figura 1. Distribución de documentos que contienen el término 'Big Data' en ProQuestResearch Library

Este cambio tecnológico y a su vez de la sociedad ha provocado un aumento de las fuentes de información y de los datos, de manera que cualquier persona, empresas u organización, puede generar información que puede ser relevante para el negocio de las empresas o gobiernos.

Por ejemplo, si un equipo directivo se encuentra en una situación de riesgo de su empresa y debe tomar una decisión, éste solicitará una investigación de

mercado a través del uso de herramientas con el fin de que dicha información les ayude a tomar la decisión más conveniente. Pero este proceso que en otro momento podría haberse llevado a cabo de manera relativamente sencilla, hoy en día se ha convertido en un proceso difícil de tratar debido al aumento constante de las cantidades de fuentes y datos, lo que ha provocado que no se puedan llevar a cabo los tratamientos de información de manera manual e incluso el software tradicional no pueda realizar en un tiempo razonable las tareas de captura y gestión de la información.

Pero además de los problemas anteriormente mencionados, nos encontramos con una nueva problemática asociada a la heterogeneidad de datos que nos podemos encontrar en la red. Para la resolución de esta problemática, sería recomendable la realización de una buena clasificación la cual nos ayudase a solucionar el problema, pero debido al continuo avance tecnológico, en el cual nos encontramos, esta clasificación puede verse afectada a lo largo del tiempo. A continuación se muestra lo que podría ser una clasificación inicial la cual nos puede ayudar como punto de partida, véase figura 2.

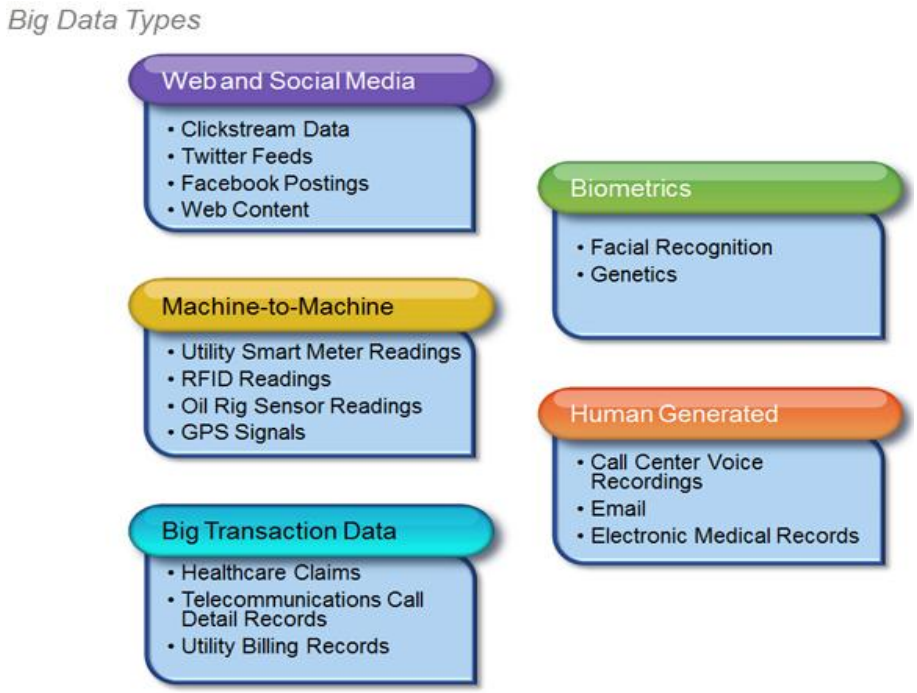

Figura 2. Clasificación de Información. Fuente IBM.

Dentro de estos cinco grupos anteriormente mencionados, centraremos nuestra atención en *Web and Social Media*. Posteriormente, comentaremos que nos ofrecen cada una exactamente.

Por tanto, localizar las fuentes, identificar información relevante dentro de la fuente y almacenar la información que generan, la cual puede tener diferentes formatos, es el primer paso de todo el proceso, el cual tiene que ser ejecutado

de manera correcta ya que el resto de fases de análisis y toma de decisiones dependen de las fuentes y datos recolectados sean correctos.

Por ello, el modelo que se propone a lo largo del trabajo se centra en la creación de un robot de conocimiento junto con los sistemas de captura de información online para fuentes hipertextuales y redes sociales, el cual identificara la información relevante dentro de una fuente, y almacena dicha información en una BBDD.

1.2. Estudio del arte

1.2.1. Big Data

Big data como hemos dicho anteriormente hace referencia al manejo de grandes volúmenes de datos de cualquier tipo. Los científicos suelen encontrar ciertas dificultades debido a la gran cantidad de datos en ciertas áreas, tales como pueden ser la meteorología, la genómica, la conectómica, las complejas simulaciones de procesos físicos...

Estos mismos problemas también se pueden encontrar en los motores de búsqueda de internet. Los *datasets*se encuentran en constante crecimiento debido en parte a la introducción de nuevos dispositivos como pueden ser los sensores inalámbricos y los dispositivos móviles (por ejemplo las VANETs), del crecimiento a lo largo de los años de los históricos de las aplicaciones (por ejemplo los logs)...

Figura 2. Big Data.

Existen una extensa variedad de tipos de datos que pueden llegar a ser capturados y analizados, una buena clasificación en el momento de su captura nos podría ayudar a entender mejor su representación (véase figura 1), aunque

están categorías podrían cambiar en el futuro para adaptarse mejor a nuestras necesidades.

- **Web and Social Media**: Incluye contenido web e información que es obtenida de las redes sociales como Facebook, Twitter, LinkedIn, etc, blogs.
- **Machine-to-Machine** (M2M): M2M se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salinidad, etc.) los cuales transmiten a través de redes alámbricas, inalámbricas o hibridas a otras aplicaciones que traducen estos eventos en información significativa.
- **Big Transaction Data**: Incluye registros de facturación, en telecomunicaciones registros detallados de las llamadas (CDR), etc. Estos datos transaccionales están disponibles en formatos tanto semi-estructurados como no estructurados.
- **Biometrics**: Información biométrica en la que se incluye huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc. En el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación.
- **Human Generated**: Las personas generamos diversas cantidades de datos como la información que guarda un call center al establecer una llamada telefónica, notas de voz, correos electrónicos, documentos electrónicos, estudios médicos, etc.

Además no es la única forma de catalogar estos datos, pues dependiendo de la forma en la cual sean almacenados, pueden catalogarse de la siguiente forma:

- **Datos estructurados (*Structured Data*):** Datos que tienen bien definidos su longitud y su formato, como las fechas, los números o las cadenas de caracteres. Se almacenan en tablas. Un ejemplo son las bases de datos relacionales y las hojas de cálculo.
- **Datos no estructurados (*Unstructured Data*):** Datos en el formato tal y como fueron recolectados, carecen de un formato específico. No se pueden almacenar dentro de una tabla ya que no se puede desgranar su información a tipos básicos de datos. Algunos ejemplos son los PDF, documentos multimedia, e-mails o documentos de texto.
- **Datos semi-estructurados (*Semistructured Data*):** Datos que no se limitan a campos determinados, pero que contiene marcadores para separar los diferentes elementos. Es una información poco regular como para ser gestionada de una forma estándar. Estos datos poseen sus propios metadatos semiestructurados[16] que describen los objetos y las relaciones entre ellos, y pueden acabar siendo aceptados por convención. Un ejemplo es el HTML, el XML o el JSON.

1.2.2. Robots de búsqueda

Definimos robot de búsqueda como aquel programa que atraviesa una estructura de hipertexto, recuperando la información para la que se ha diseñado, por ejemplo, los grandes motores de búsqueda de la web se alimentan de un robot de búsqueda diseñado para la recolección de enlaces.

Los robots son usualmente llamados “Web Wanderers”, “web Crawlers” o, como son más comúnmente conocidos, “Spiders” (arañas de búsqueda) y suelen visitar los sitios web y extraer la información deseada que están incluidos dentro de estas web.

Su forma de ejecución depende de cómo estén diseñados, en general comienzan a trabajar desde una lista de URL’s. Esto les da un punto de partida para comenzar a seleccionar url’s que han de visitar, analizarlas y usarlas como recurso para incluirlas dentro de su base de datos.

1.2.3. Knowbot

Un knowbot es un tipo de robot que recoge datos o información mediante la recopilación automática de cierta información específica previamente de los sitios web.

KNOWBOT es un acrónimo de la palabra inglesa Knowledge-Based Object Technology. Este término, utilizado ya en 1990, describe objetos basados en computadoras desarrolladas para recoger y almacenar información específica, con el fin de utilizar dicha información para realizar una tarea específica, y para permitir compartir esa información con otros objetos o procesos. Un uso temprano de knowbots era proporcionar un asistente automatizado para los usuarios para completar tareas detalladas redundantes sin necesidad de capacitar al usuario en la tecnología informática.

2. Nuestra Propuesta

El modelo que se propone en este trabajo es el desarrollo de un sistema que se encargue de la búsqueda, selección y recopilación de información en internet dada una empresa.

Para la búsqueda de dicha información usaremos distintos métodos dependiendo de la procedencia de la información:

- Búsqueda en tiempo real.
- Búsqueda programada.

Una vez realizada la búsqueda, se procederá a la selección y recopilación de la información útil, separándola y catalogándola por apartados, procedencia, tipo..., para su posterior guardado.

Para la mayoría de los procesos realizados se utilizaran expresiones regulares simples y complejas, que a partir de ahora los denominaremos como patrones. Su uso será explicado posteriormente.

Una vez que la información se encuentra recolectada en la base de datos, esta puede ser visualizada a través de una pequeña interfaz web.

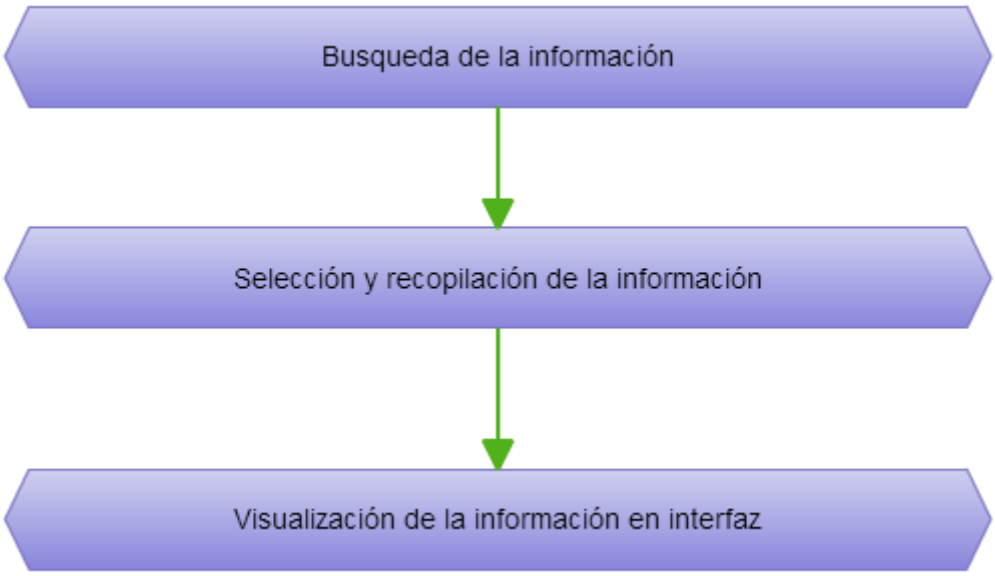

Figura 4. Esquema del sistema

Se realizó un estudio de diferentes fuentes hipertextuales y redes sociales desde donde poder obtener información, del que finalmente nos quedaremos con tres fuentes: Twitter, prensa digital y foros.

Dependiendo de su procedencia se aplicaran los distintos módulos de tal manera que se puedan adaptar lo máximo posible a las necesidades fuente.

Las entidades que van a resultar objetivo de nuestro proyecto serán del sector de la telecomunicación, pero nos centraremos únicamente en las empresas, sino que también buscaremos a sus respectivos presidentes o consejeros responsables de las empresas en España. Véase tabla 1.

Tabla 1. Empresas y directivos

Empresa	Persona
Telefónica	César Alierta
Vodafone	Antonio Coímbra
Orange	Luis Alberto Salazar
Xfera Móviles (Yoigo)	Eduardo Taulet

2.1. Prensa Digital

La mayoría de los diarios digitales poseen fuentes web, las cuales se utilizan para difundir información actualizada frecuentemente a usuarios que se han suscrito a la fuente de contenidos. El formato permite distribuir contenidos sin necesidad de un navegador.

Existen muchos formatos para las fuentes web, entre el que destaca el formato RSS (Really Simple Syndication) el cual es parte de la familia de los formatos XML, desarrollado específicamente para todo tipo de sitios que se actualicen con frecuencia y por medio del cual se puede compartir la información y usarla en otros sitios web o programas. A esto se le conoce como redifusión web o sindicación web (una traducción incorrecta, pero de uso muy común).

Figura 5. Ejemplo de RSS

Por tanto, nosotros usaremos las fuentes web como punto inicial de las búsquedas de las distintas noticias.

Para la selección de las noticias, nuestro sistema se encargara de filtrar las noticias de aquellas empresas que buscamos de aquellas que no nos son útiles. Para este filtrado utilizamos dos sistemas de patrones distintos:

- **Sinónimos**: son aquellas palabras o conjunto de palabras que hacen referencia a la empresa que buscamos.
- **Anti-Sinónimos**: son aquellos conjuntos de palabras que nos indican que el sinónimo que dábamos por valido, realmente se está utilizando con un significado distinto de la palabra.

Tabla 2. Ejemplo de sinónimo y anti-sinónimo

Empresa: Ono	Sinónimo: Ono	Anti-Sinónimo: Yoko Ono
Noticia 1: Vodafone compra Ono.	Sinónimo valido	
Noticia 2: El increíble romance de Yoko Ono.	Sinónimo invalido	

Como podemos ver en la tabla 2, en ambas expresiones aparece la expresión Ono, por tanto si solo mirásemos el sinónimo, ambas serian noticias validas de tal manera que nuestro filtrado no sería del todo correcto y fiable, pero al aplicar el sistema de Anti-Sinónimos, podemos descartar la 2º noticia como valida de tal manera que haríamos un filtrado más correcto.

Una vez que hemos terminado con el sistema de filtrado y la notica es válida, esta pasa a una cola para la extracción de datos.

En este modulo procedemos a obtener los diferentes parámetros que nos ofrece la notica, como pueden ser el titulo, la cabecera, la url, el cuerpo de la noticia, el autor…

Realizaremos un recopilador que lleve a cabo una búsqueda en el HTML basado en patrones. Con estos patrones se buscan los parámetros deseados, en la figura 6 podemos ver algunos de ellos.

Estos patrones utilizados pueden ser:

- **Simples**: que con una sola expresión regular podemos capturar todos los parámetros que necesitemos,
- **Compuestas**: que se necesite de varias expresiones regulares para poder obtener un parámetro (Tabla 3).

Tabla 3. Patrones necesarios para captura de un cuerpo.

Parámetro	Patrones
Inicio Noticia	`"<div id=\"tamano\">"`
Cuerpo Noticia	`"<h3>(.*)</h3>"`
Foto en noticia	`"<div class=\"sumario derecha\">"`
Fin Noticia	`"<script type=\"text/javascript\">"`

Figura 6. Algunos de los parámetros de una noticia.

El procedimiento que se ha seguido para la elaboración de estos patrones se explica a continuación:

I. Partiendo de los un conjunto de noticias seleccionadas aleatoriamente, se definirán un conjunto de patrones.
II. Haciendo uso de los patrones ya existentes, se añadirán los patrones complementarios que no se encuentren definidos.
III. Finalmente se harán combinaciones para obtener patrones que no se hayan determinado hasta este momento.

Una vez recopilado los parámetros, se inicia el guardado en la base de datos.

2.2. Foros

Los foros están estructurados como árboles (ver figura 7), por ello usaremos esta estructura a nuestro favor para realizar las búsquedas, dándole por punto inicial de lectura el punto más alto que podamos seleccionar, en nuestro caso, la página inicial del foro y poder ir descendiendo por las distintas secciones del foro hasta llegar a los temas de discusión de los usuarios.

Para poder llevar a cabo esta búsqueda, utilizaremos patrones obtenidos a través del estudio del foro seleccionado. Una vez que se llega al nivel de post (respuesta de los usuarios en los temas de discusión), se inicia el filtrado con el módulo de selección.

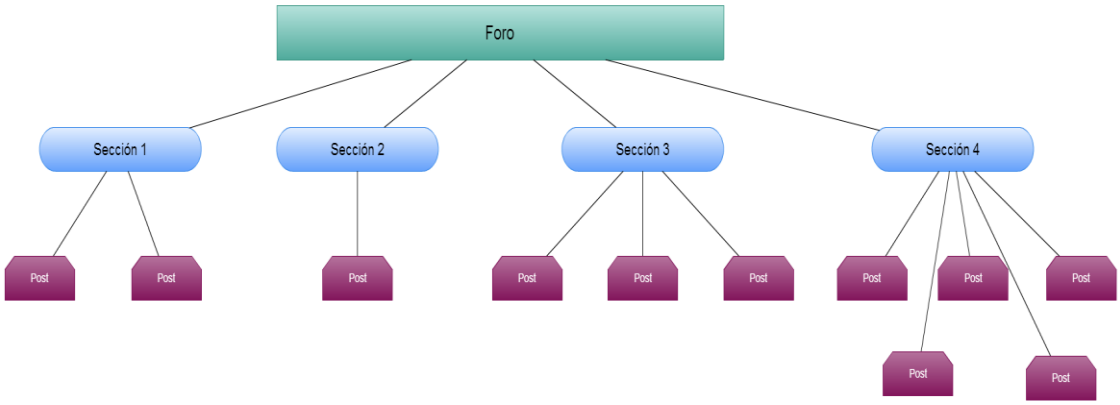

Figura 7. Estructura interna de un foro.

La parte de selección y filtrado funciona igual que explicada anteriormente en la sección 2.1. El sistema de filtros se aplicara en el titulo de los temas de discusión de los usuarios.

Finalmente para la recolección de los parámetros establecidos procedemos a obtener los diferentes parámetros que nos ofrece los temas de discusión y post de los usuarios, para obtener estos datos los separaremos en 2 entidades distintas:

- **Tema**: que posee el post inicial que inicia la discusión, el titulodel tema, fecha, hora, la url de la primera y de la última página escrita del tema de discusión,…
- **Post**: son todas las respuestas al tema, por tanto aquí guardaremos cual es el tema al que responde, si es respuesta a otra respuesta dentro del mismo tema, la fecha y hora,…

Para poder obtener estos parámetros haremos uso de patrones obtenidos con un estudio similar a la explicada 2.1, aplicado a la estructura del foro que estemos estudiando en cuestión.

2.3. Twitter

Este módulo es el que más dificultades presenta, pues el tipo de escritura que podemos encontrar en la red social Twitter, es totalmente abierta. Esto conlleva que para realizar la búsqueda debemos crear expresiones complejas lo más amplias posibles para poder capturar la mayor cantidad de información posible.

Por ejemplo, para la empresa Telefónica tendríamos las siguientes opciones:

- Teléfonica
- Telefonica
- telefónica
- telefonica

Estas cuatro combinaciones, entre otras muchas otras, son posibles encontrarlas en Twitter haciendo referencia a la empresa Telefónica, por tanto es posible que incluso después del estudio realizado en Twitter para poder encontrar todos los sinónimos posibles, haya algunos que no hayamos incluido.

Una vez que hemos conseguido las expresiones regulares complejas lo más completas posibles en el sistema de búsqueda, el sistema de selección resulta menos complejo, pues los resultados de la búsqueda ya nos darán la información filtrada, solo sería necesario la aplicación de anti-sinónimos para darla por válida.

Una vez dada por validada la información obtenida, empezamos con el sistema de recopilación de datos, el cual se encargara de obtener los parámetros que nosotros hemos establecidos, como pueden ser el cuerpo del tweet, la fecha y hora, la localización,…

3. Implementación

3.1. Base de datos

En este epígrafe se hará un breve resumen para presentar el que será nuestro modelo Entidad/Relación a lo largo del trabajo.

El sistema de gestión de base de datos utilizado en este trabajo será Microsoft SQL Server 2012. Sus lenguajes para consultas son T-SQL y ANSI SQL. Microsoft SQL Server constituye la alternativa de Microsoft a otros potentes

sistemas gestores de bases de datos como pueden Oracle, MySQL y PostgreSQL.

3.1.1. Lectura de Información y Entidades

Los elementos básicos de los procesos de lectura quedaran almacenados en la tabla información, y dependiendo de su precedencia (red social, prensa digital o foro), se almacenara en alguna de sus subcategorías, tal y como se muestra en la figura 8.

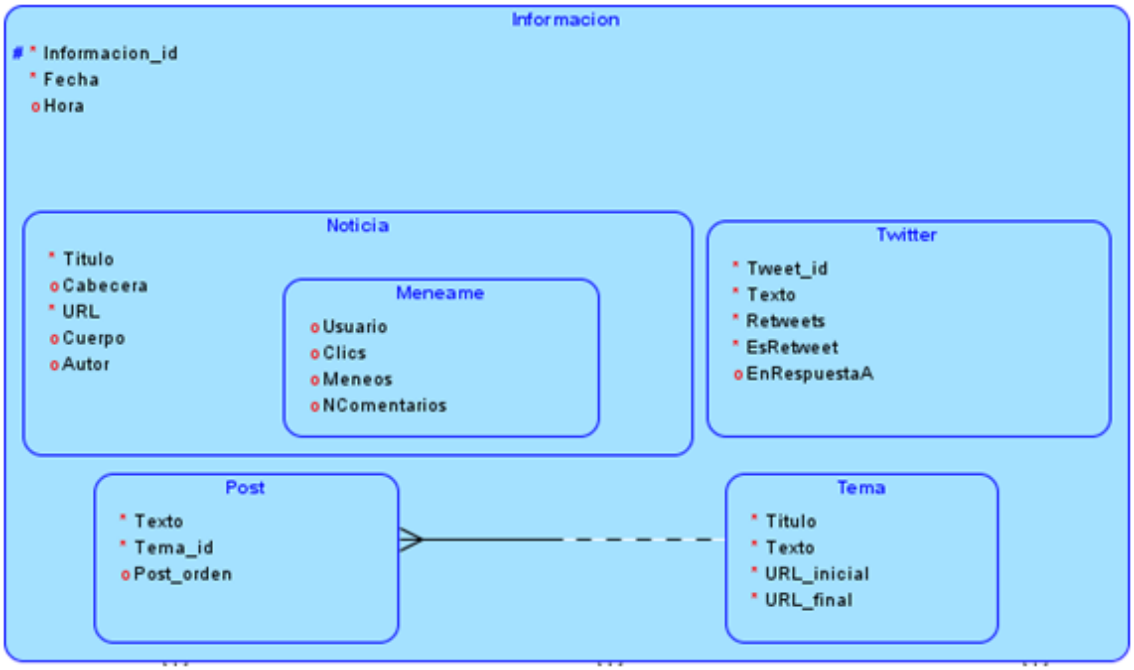

Figura 8. Entidad información.

La única excepción se trata de la web Menéame, la cual es una combinación entre red social y prensa digital, pues en ella los distintos usuarios que están registrados añaden noticias para que el resto de usuarios las lean, comenten y compartan. Tras estudiarla se decidió incluirla como una subclase de la entidad Noticia ya que era a la que más se asemejaba.

La información recopilada también posee algunos parámetros extras como pueden ser la procedencia, el idioma y el origen de la información (ver figura 9). En caso de que el origen fuese la prensa digital, también tendremos guardado cual fue el RSS de la fuente.

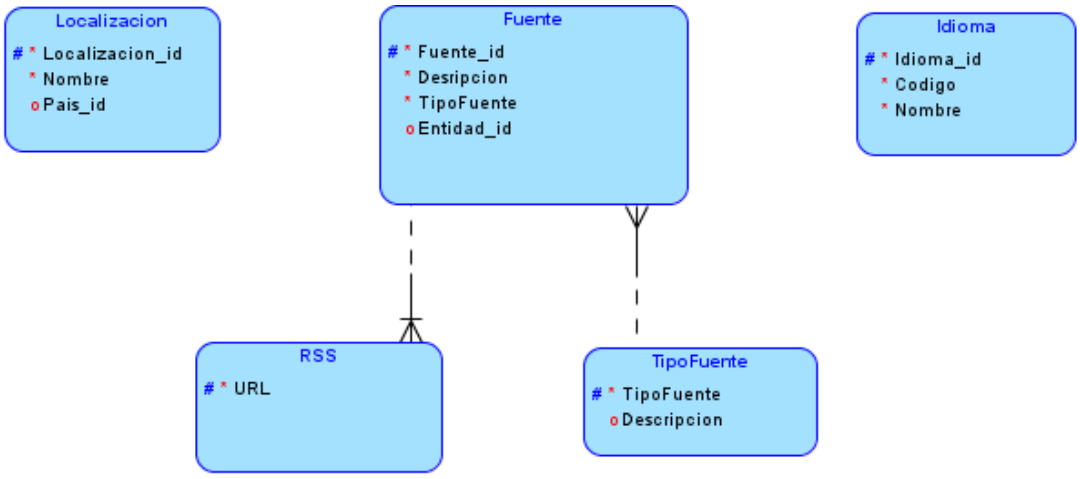

Figura 9. Entidades relacionadas con información.

La información recopilada hará referencia hará referencia a una entidad (ver figura 10). Esta entidad puede ser una empresa o una persona.

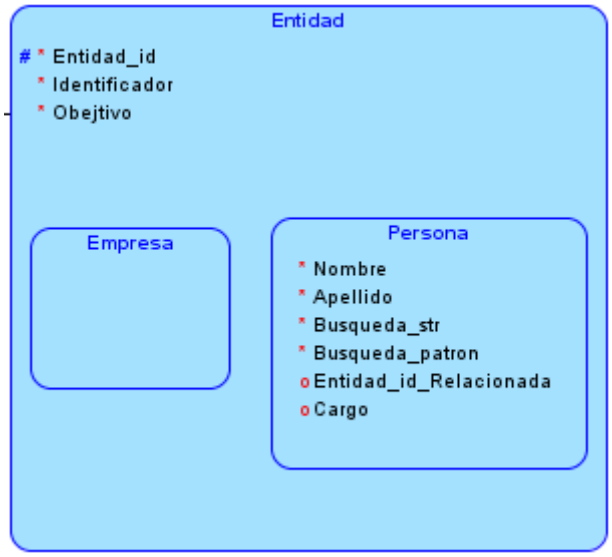

Figura 10. Entidad entidad.

Como hemos explicado anteriormente, la información no siempre aparece con el nombre exacto de la entidad, por tanto la Información y entidad están relacionadas por los sinónimos de la entidad (ver figura 11).

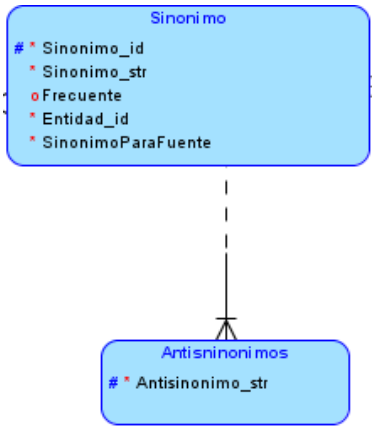

Figura 11. Entidad Sinónimo y Anti-Sinónimo.

3.1.2. Uso de patrones

Para la captura de información, utilizamos patrones (ver figura 12). Dependiendo del patrón pueden ser simples o compuestos y los compuestos se deben aplicar en un orden concreto. Esto se recopila en la base de datos.

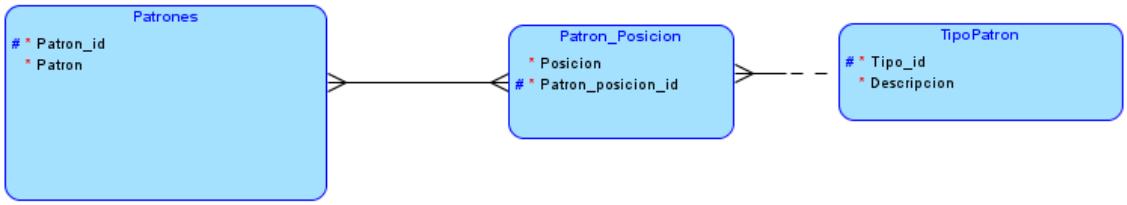

Figura 12. Entidad Patrones.

3.1.3. Usuarios del sistema

El sistema está pensado para ser visualizado en una aplicación web, por tanto necesitaremos un sistema de usuarios y perfiles para poder controlar la visión de cada uno, véase figura 13.

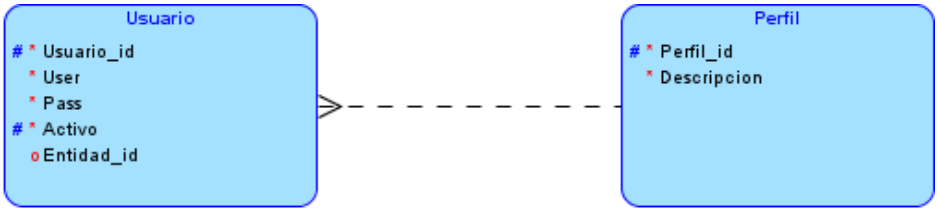

Figura 13. Entidad Usuario y Perfil.

Si unimos todas las entidades explicadas aquí, el sistema queda de la siguiente manera (Figura 14).

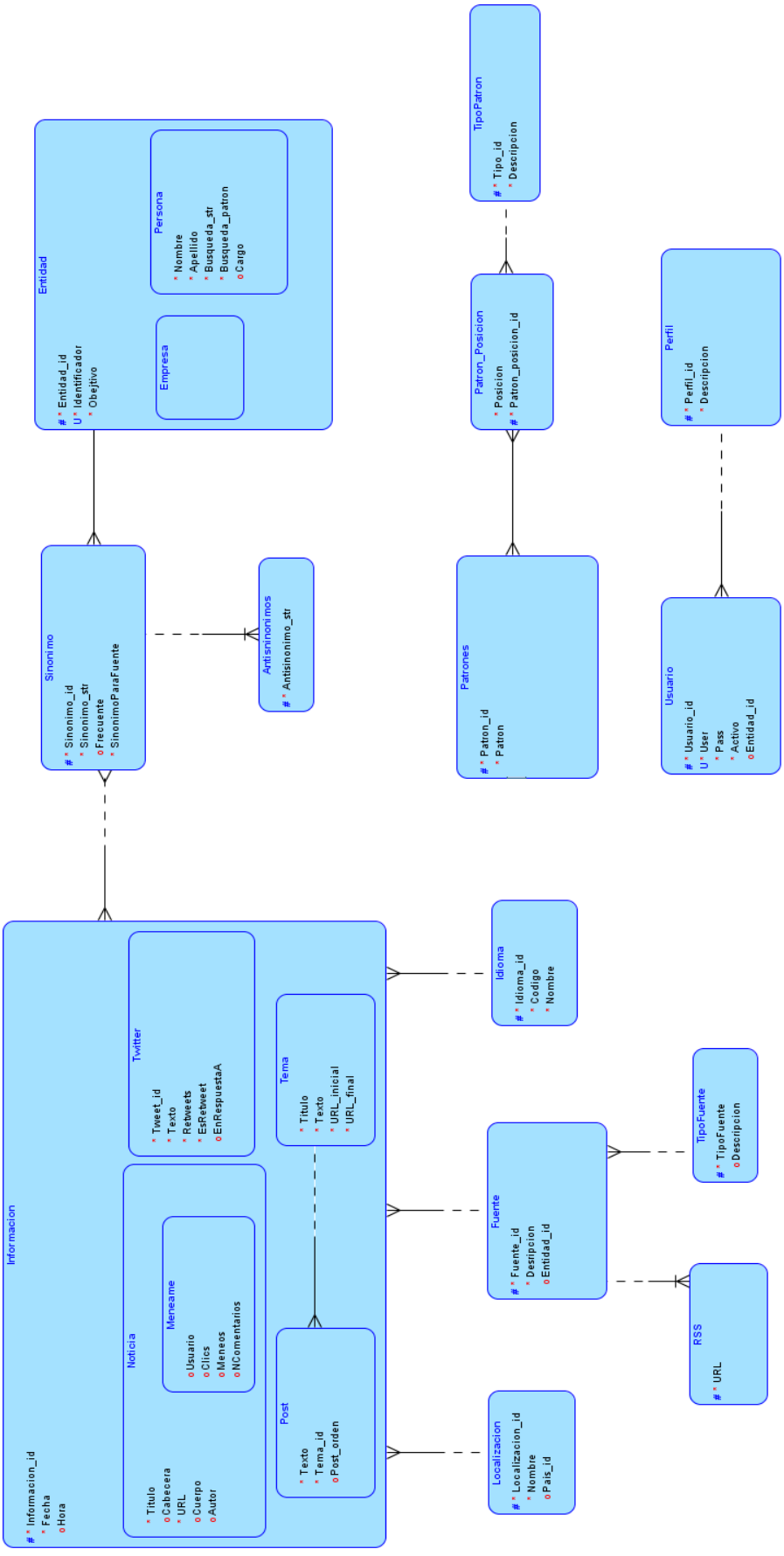

Figura 14. Modelo Entidad-Relación completo.

3.2. Herramientas de desarrollo

3.2.1. Eclipse

El entorno de desarrollo escogido para la realización del código de programación para la lectura ha sido Eclipse.

Este entorno nos permite programar usando distintos lenguajes de programación como pueden ser C, C++, Java, HTML, XML,… Para nuestro proyecto hemos elegido el lenguaje JAVA, por su gran uso y al ser el lenguaje de programación que más se ha trabajada lo largo de la carrera, por lo cual nos resulta más familiar.

Eclipse es un programa informático compuesto por un conjunto de herramientas de programación de código abierto que puede ser descargado desde su página web https://www.eclipse.org/downloads/

3.2.2. Hibernate

Hibernate es la librería que hemos elegido para poder operar con la base de datos usando código JAVA.

Hibernate es una herramienta de Mapeo objeto-relacional (ORM) para la plataforma Java (y disponible también para .Net con el nombre de NHibernate) que facilita el mapeo de atributos entre una base de datos relacional tradicional y el modelo de objetos de una aplicación, mediante archivos declarativos (XML) o anotaciones en los beans de las entidades que permiten establecer estas relaciones.

Hibernate es software libre, distribuido bajo los términos de la licencia GNU LGPL, el cual puede ser descargado desde el siguiente enlace http://hibernate.org/orm/downloads/.

3.3. Lectura Twitter

Esta sección contendrá una breve introducción a los conceptos más comunes a la hora de hablar de Twitter, así como un análisis de las diferentes alternativas existentes para leer información de dicha red. En la conclusión del epígrafe se mostrará la solución adoptada de entre las planteadas.

3.3.1. Conceptos Básicos

Términos comunes en el uso de Twitter:

- **Tweets**: Los tweets son las unidades de información que se emiten a través de Twitter. Estas unidades poseen una singular característica que actúa de limitación; su longitud no puede ser superior a 140 caracteres.
- **Retweets**: Mecanismo mediante el cual un usuario tiene la posibilidad de dar mayor difusión a un tweet, ya que mediante esta acción el tweet se estaría compartiendo con sus seguidores, que visualizarían el mismo en su timeline.
- **Menciones**: Este concepto hace alusión a la funcionalidad que permite citar a un usuario en un mensaje. Para ello, es necesario usar el símbolo @ seguido del nombre del usuario.
- **Respuesta**: Un tweet escrito por un usuario de Twitter puede ser rebatido por otro miembro de la comunidad. De esta forma el mensaje aparecerá en el timeline de los usuarios comunes a ambos.
- **MensajeDirecto**: Es un mensaje privado entre dos usuarios de Twitter. Para poder enviar un mensaje privado a un usuario es necesario que ambos estén siguiéndose mutuamente.
- **Hashtag**: Un hashtag queda definido por la cadena de texto que sigue al símbolo #. Su fama recae en la utilidad a la hora de agrupar conversaciones en torno a un mismo tema.
- **Followers**: Usuarios que siguen a un determinado usuario. Todos los mensajes escritos por un usuario, quedarán reflejados en el timeline de sus followers.
- **Following**: Es el concepto inverso a Followers. Este concepto hace referencia a las personas a las que un miembro de Twitter sigue. Los mensajes que emiten quedarán reflejados en su timeline.
- **Timeline**: Su traducción es línea de tiempo y en ella aparecen todos los Tweets de las personas a las que un usuario está siguiendo además de los que el propio miembro escribe.
- **TrendingTopic**: Son los temas más hablados del momento, los temas de actualidad.

3.3.2. Opciones para lectura de Twitter

Twitter provee diferentes APIs para facilitar el acceso a sus datos. De todas ellas, para nuestro objetivo de recabar tweets, nos encontramos con dos que cumplen con los requisitos:

- REST API

- API STREAMING

Cada una de ellas posee una serie de características y limitaciones, las cuales vamos a exponer a continuación.

3.3.2.1. REST API

Proporciona gran cantidad de interfaces que engloban las distintas funcionalidades que ofrece Twitter. Entre estas interfaces se encuentran las siguientes: Timeline, Tweets, Search, Streaming, Direct Messages, Friends & Followers, Users, Suggested Users, Favorites, Lists, Saved Searches, Places & Geo, Trends, Spam Reporting, OAuth, Help.

De todas las interfaces anteriormente citadas, la que más se adecua a nuestro objetivo es la interfaz de búsqueda 'Search'. A continuación se muestra un listado extraído de la documentación oficial de Twitter con los distintos parámetros que acepta la interfaz.

Tabla 4. Parámetros de REST API

Parameters	
Q Required	A UTF-8, URL-encoded search query of 500 characters maximum, including operators. Queries may additionally be limited by complexity.
Geocode Optional	Returns tweets by users located within a given radius of the given latitude/longitude. The location is preferentially taking from the Geotagging API, but will fall back to their Twitter profile. The parameter value is specified by "latitude, longitude, radius", where radius units must be specified as either "mi" (miles) or "km" (kilometers). Note that you cannot use the near operator via the API to geocode arbitrary locations; however you can use this geocode parameter to search near geocodes directly. A maximum of 1,000 distinct "sub-regions" will be considered when using the radius modifier.
Lang Optional	Restricts tweets to the given language, given by an ISO 639-1 code. Language detection is best-effort.
Locale Optional	Specify the language of the query you are sending (only ja is currently effective). This is intended for language-specific consumers and the default should work in the majority of cases.
result_type optional	Optional. Specifies what type of search results you would prefer to receive. The current default is "mixed." Valid values include: * mixed: Include both popular and real time results in the

	response. * recent: return only the most recent results in the response * popular: return only the most popular results in the response.
Count Optional	The number of tweets to return per page, up to a maximum of 100. Defaults to 15. This was formerly the "rpp" parameter in the old Search API.
Until optional	Returns tweets generated before the given date. Date should be formatted as YYYY-MM-DD. **Keep in mind that the search index may not go back as far as the date you specify here.**
since_id optional	Returns results with an ID greater than (that is, more recent than) the specified ID. There are limits to the number of Tweets which can be accessed through the API. If the limit of Tweets has occurred since the since_id, the since_id will be forced to the oldest ID available.
max_id optional	Returns results with an ID less than (that is, older than) or equal to the specified ID.
include_entities optional	The entities node will be disincluded when set to false.
Callback Optional	If supplied, the response will use the JSONP format with a callback of the given name. The usefulness of this parameter is somewhat diminished by the requirement of authentication for requests to this endpoint.

En la sección para desarrolladores de Twitter se encuentra una entrada que resume las buenas prácticas en el uso de dicha interfaz.

Las limitaciones más importantes asociadas al uso de esta API son las siguientes:

- La ventana temporal de consulta estará limitada (entre 6-9 días anteriores).
- Dependiendo de la tipología de la clave usada (por usuario o por aplicación) existirá un número máximo de peticiones en ventanas temporales de 15 minutos (180,450 respectivamente).

3.3.2.2. API STREAMING

El conjunto de APIsStreaming que proporciona Twitter posibilita el acceso de baja latencia al Stream global de Tweets.

Twitter suministra una serie de streams diferenciados, cada uno de ellos con un propósito distinto. A continuación, se listan los streams existentes:

- PublicStream
- UserStream
- SiteStream

Para la elaboración de este trabajo, el Stream que a priori parece más adecuado sería el de usuarios. En la documentación oficial de Twitter se recoge la definición del mismo: permite el acceso al Stream de los datos públicos que fluyen a través de Twitter. Se recomienda su uso para seguir usuarios o temas específicos, así como para minería de datos.

Las limitaciones más destacadas de su uso son las siguientes:

- El número máximo de tweets recibidos es equivalente a una pequeña fracción del volumen total de tweets generados en un instante determinado

3.3.2.3. Medidas tomadas

Teniendo en cuenta el análisis realizado y nuestras necesidades para llevar a cabo el proyecto, la API escogida será Streaming API.

En la propia página de Twitter se puede encontrar la siguiente recomendación:

> *"Si tu aplicación requiere de repetidas peticiones a la Search API, deberías considerar el uso de Streaming API"*

Aclarada la solución que se adoptará, es importante detallar que se hará uso de la librería Twitter4J. Se trata de una librería no-oficial para Twitter API, que permite integrar de formar sencilla los servicios de Twitter en aplicaciones Java.

3.4. Lectura Noticias

Tal y como hemos explicado en el sección 2.1, para la lectura de prensa digital haremos uso de las fuentes Web, por ello usaremos la librería ROME.

Rome es una librería para Java usada para leer y generar contenido en formato RSS, esta librería es un software de código abierto bajo la licencia de Apache 2.0.

Aun así, habrá ciertos periódicos que debemos leer completamente con patrones debido a que no disponen de RSS públicos

Los periódicos seleccionados para este proyecto con sistema de RSS público son:

- ABC
- Agencia EFE
- Cinco Días
- El Economista
- El Mundo
- El País
- Europa Press
- La Razón
- La vanguardia
- Libertad Digital
- Menéame

Estas fuentes disponen de un subprograma común al que se le suministra los RSS que previamente hemos añadido al sistema, el cual los lee y rellana los campos de la base de datos con la información leída.

Además cada periódico tiene un sistema propio con un sistema de patrones que recolecta la información de la noticia que no aparece en los RSS a través de la URL almacenada en la base de datos.

Mientras que los periódicos seleccionados sin RSS público son:

- Colpisa
- Reuters

Las lecturas de estos periódicos se realizan completamente con el sistema de patrones. Para ello se les da un punto de origen donde aparecen las noticias del día, y desde ese punto empieza a leer todos los datos que son necesarios para el sistema.

3.5. Lectura Foros

Como las entidades seleccionadas para las búsquedas pertenecen al ámbito de las telecomunicaciones hemos buscado un foro centrado en dicho ámbito, por ello el foro seleccionado ha sido http://www.adslzone.net/foros.html

Al no haberse encontrado ningún formato como las fuentes web que aparecen en la prensa digital, hemos tenido que optar por un método más personalizado, realizar un estudio sobre el código fuente de la página adslzone para poder crear nuestro propio lector basado en patrones de búsqueda.

Como hemos explicado anteriormente en el apartado 2.2, los foros tienen una estructura de árbol, por tanto nuestra búsqueda la iniciaremos en la página inicial del foro, la cual es la raíz para posteriormente ir descendiendo por sus hojas.

Tal y como podemos ver en la figura 15, nos encontramos en la raíz y en la cual se ven distintas secciones, como la remarcada en naranja que hace se tratan todos los temas relacionados con el adsl.

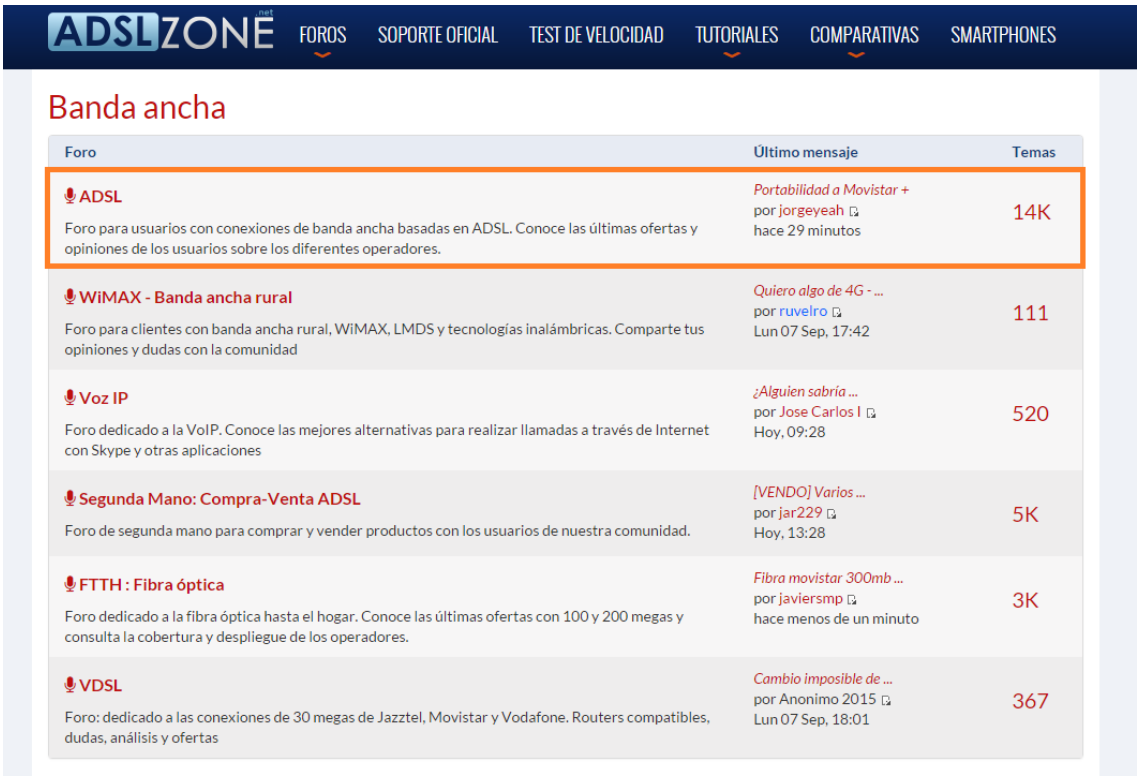

Figura 15. Ejemplo de secciones en el foro

Una vez que hemos accedido a esta sección a través de los patrones establecidos, nos encontramos en el siguiente nivel del árbol, en el cual podemos ver todos los temas que contiene esta sección (ver figura 16).

Figura 16. Ejemplo de temas.

3.6. Lector compuesto

Una vez creados todos los lectores, es momento de unirlos para crear nuestro sistema de lectura compuesto. Para ello realizaremos 2 hebras que trabajen en paralelo.

Una de ellas contendrá el lector de twitter, el cual al ser un streaming, nos dará información continua y en tiempo real, lo que nos impide leer información anterior, por tanto esta hebra nunca se dormirá y estará suministrando información de manera constante a la base de datos.

Otro hebra contendrá los lectores de foros y prensa digital, como estas dos webs mantienen sus datos (la prensa digital durante un periodo grande de tiempo y los foros de manera constante), esta hebra se ejecutara cada X horas realizando la lectura y luego durmiéndose. Para la realización de este proyecto se estableció una frecuencia de lectura de cada 6 horas, es decir, 4 lecturas diarias para la prensa digital y cada 12 horas para el sistema de foros.

3.7. Interfaz

Una vez que ya tenemos toda la información recopilada, necesitamos un sistema para poder llevar a cabo la visualización de la misma. Para ello se ha creado una pequeña página web para fara que todo el mundo, independientemente de su conocimiento de informática, pueda tener acceso a ella.

Esta página dispondrá de una página inicial de login de usuario (ver figura 17), en la cual todo aquel que tenga su cuenta y pass activa, podrá acceder para ver la información relacionada con su cuenta, es decir, si un usuario con acceso a la información de Vodafone, únicamente podrá ver noticas, tweets o post relacionados con la empresa Vodafone.

Una vez que ya nos hemos logueado satisfactoriamente, nos aparecerá un menú en el cual podremos elegir el rango de fecha en el que queremos buscar, que tipo de información queremos visualizar (Twitter, Noticias o Foros). Con estos campos seleccionados, hacemos uso del botón actualizar que aparece en la página (ver Figuras 18 y 19).

Si hay información en el rango de fechas se mostraran por pantalla, y si además el rango es suficientemente amplio, podemos seleccionar una grafica para que nos muestre la cantidad de información recolectada por día (ver Figura 20).

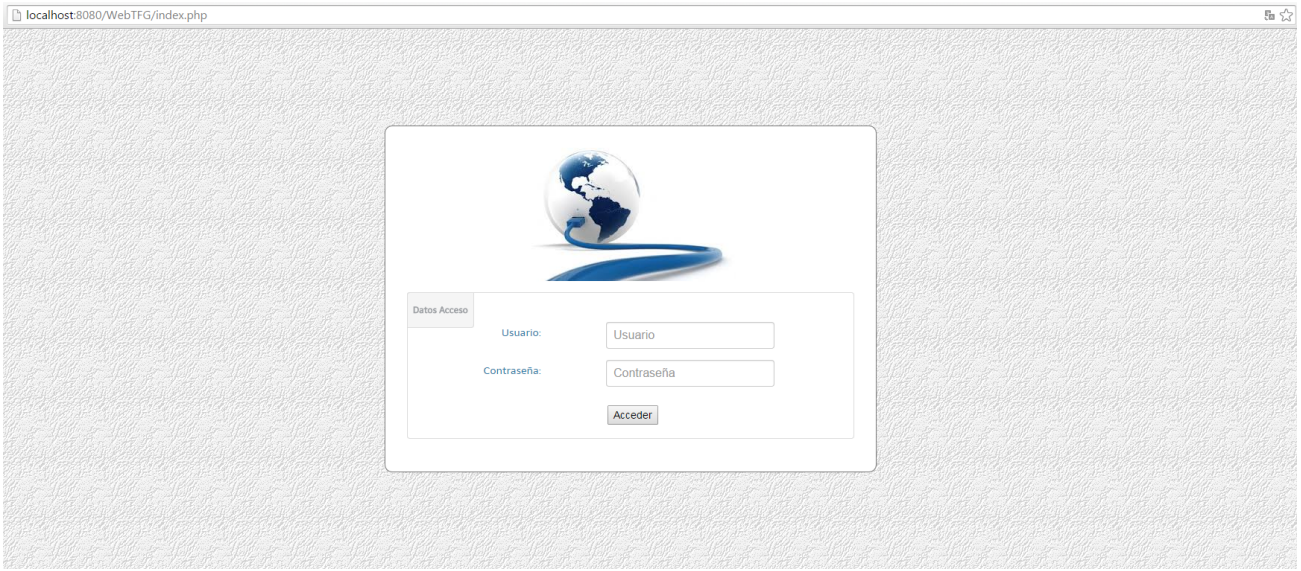

Figura 17. Página de Login.

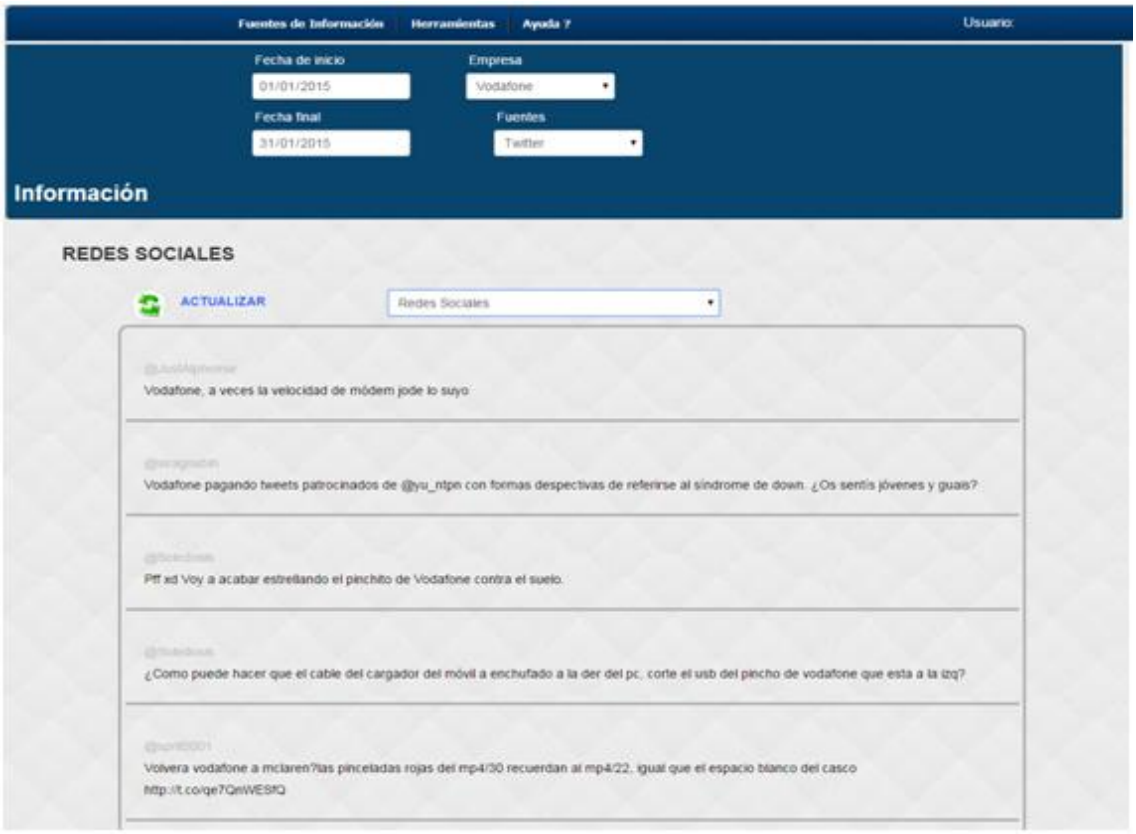

Figura 18. Twitter Vodafone

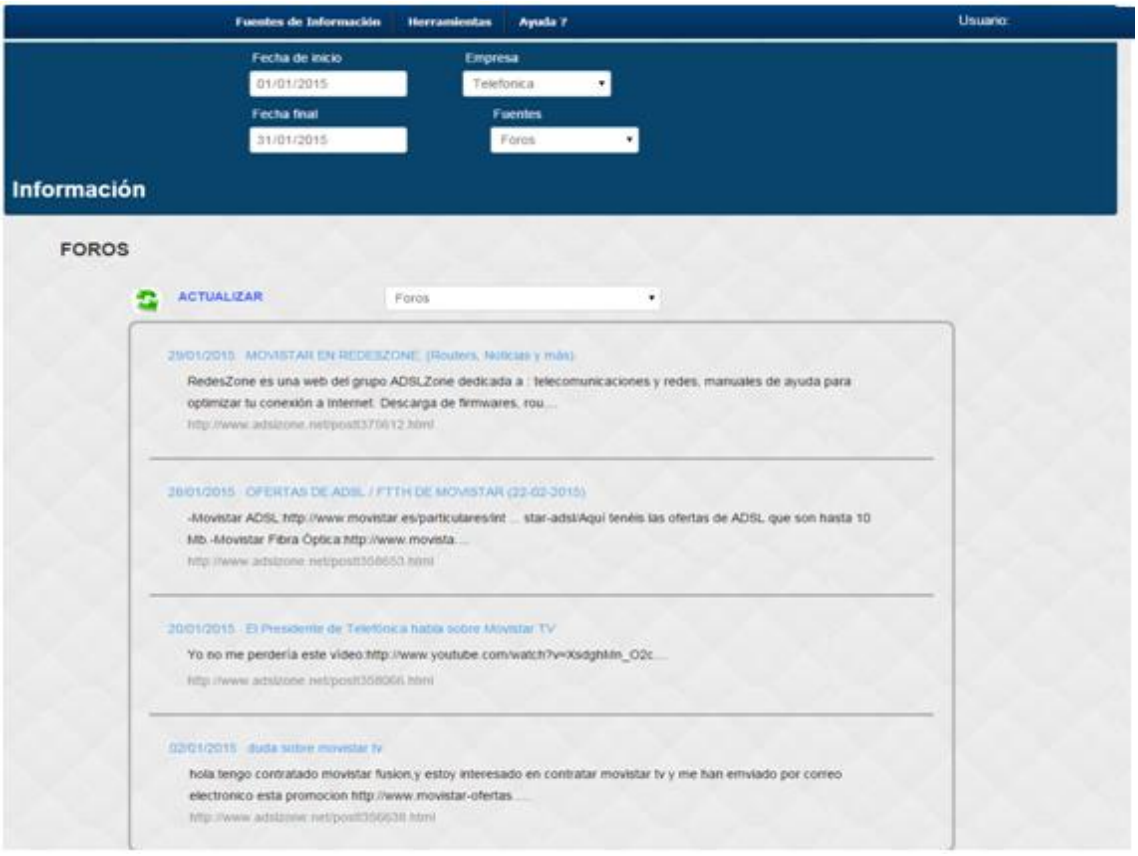

Figura 19. Foros Telefónica.

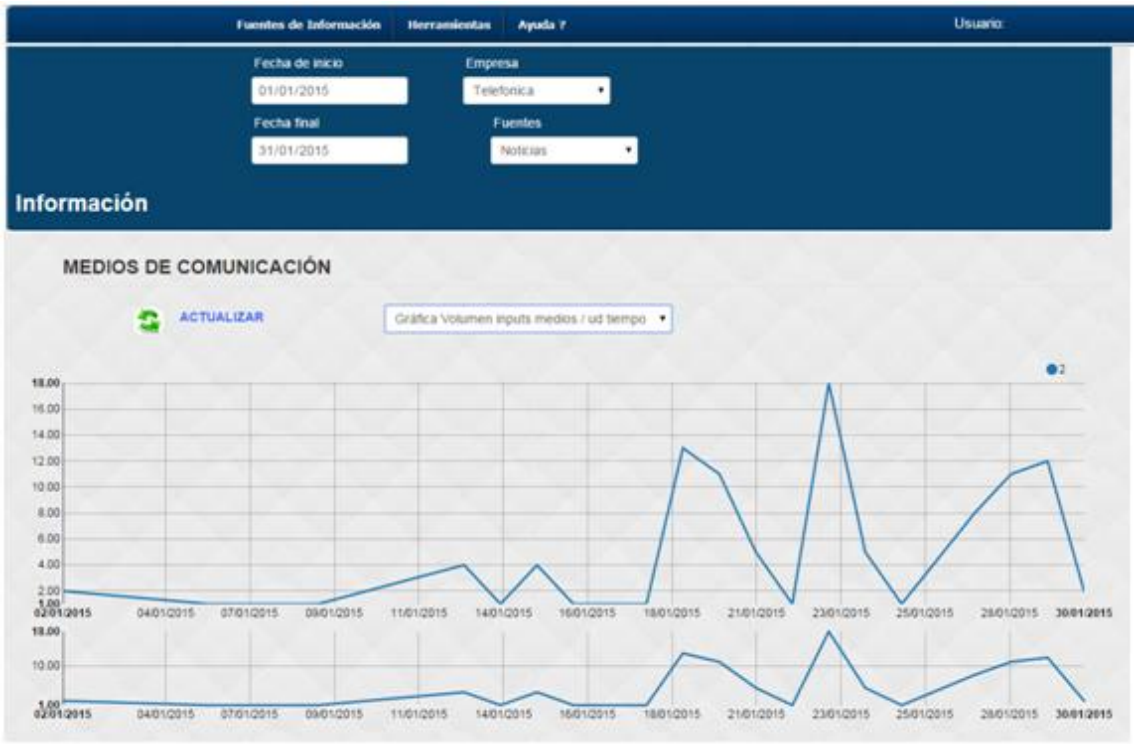

Figura 20. Grafica de noticias por día para Telefónica.

Para la creación de esta página web se ha utilizado lenguaje HTML, PHP, Javascript y se ha ejecutado en un servidor Wamp.

4. Conclusiones

Las conclusiones que se han obtenido a lo largo de la realización de este trabajo final de grado, se pueden clasificar en tres apartados: en primer lugar, el trabajo desarrollado; en segundo lugar, el aprendizaje personal que ha supuesto la realización de este trabajo; y finalmente, los trabajo futuros.

Trabajo Desarrollado

En este trabajo se ha desarrollado un robot de conocimiento para la extracción de información de los Social Media y Web, centrándonos principalmente en las redes sociales y fuentes hipertextuales. Se han desarrollado distintos buscadores y sistemas de lectura aplicando diferentes técnicas.

Para las redes sociales (Twitter), se ha creado un buscador en tiempo real basado en streaming, pues era la opción que mejor se adaptaba a nuestras necesidades de búsqueda, el cual nos suministra datos de manera abundante y constante, para que posteriormente nuestro sistema pueda almacenar la información de utilidad.

Para las fuentes hipertextuales, se han creado dos buscadores programables. El primero de ellos para prensa digital diseñando con un buscador a partir de un lector de fuentes web el cual permite el acceso a los datos, y un sistema de patrones el cual filtra y recolecta la información útil para su almacenamiento.

Y otro para foros, el cual se ha diseñado con un robot de búsqueda creado a partir de un sistema de patrones que recorre toda la estructura. Estos patrones son utilizados tanto para realizar la búsqueda de datos a través del foro como para la recolección de los mismos.

Finalmente, se ha diseñado una interfaz web para interactuar con el usuario, que permite visualizar de manera sencilla y amigable todos los datos recolectados por los lectores de búsqueda.

El sistema ha estado en ejecución los meses de Enero, Febrero y Marzo del año 2015, del cual hemos obtenido una cantidad total de 246631 tweets, 561 noticias, 9288 post, y 126 temáticas durante estos meses.

A lo largo del trabajo se ha podido comprobar que el método realizado que en la web y redes sociales podemos encontrar una gran fuente información creada, leída y compartida por los propios usuarios que puede ser de gran utilidad.

Proceso de aprendizaje

La realización de este trabajo ha supuesto la adquisición y afianzamiento de conceptos algunos de los cuales han sido adquiridos durante los años de estudio en el grado. De manera pormenorizada pero sin extenderme, se pueden resumir en:

- Poner en práctica los conocimientos adquiridos en la carrera, como por ejemplo, el estudio, modelaje y realización de nuestra propia base de datos, planificación de un proyecto,…
- Introducirme en un nuevo campo como es el Big Data.
- El desarrollo de tres sistemas de lecturas online en paralelo.

Futuros Trabajos

Los trabajo futuros que se pueden derivar, entre otros, del presente trabajo son:

- Diseño de una araña de búsqueda que permita la localización de URLs, de forma que el proceso pueda completarse con la localización de fuentes, localización y extracción de información y finalmente almacenamiento de los datos obtenidos.

- Diseño de un sistema de análisis de sentimiento, que tenga en cuenta tanto la polaridad y la intensidad de las informaciones que se emiten sobre las empresas.

5. Bibliografía

Una de las fuentes principales de bibliografía es la web de Hibernate. De donde he sacado toda la información para aprender a manejar la librería (http://hibernate.org/orm/documentation/).

También he usado la bibliografía de Java para despejar dudas referentes al código (http://download.oracle.com/javase/6/docs/api/overview-summary.html).

Bibliografías que me han sido útiles a la hora de desarrollar el TFG:

[1] Sergio A. Rojas (2012). Towards automatic recognition of irregular, short-open answers in Fill-in-theblank tests

[2] Zheng Kai, Qiaozhu Mei & David Hanauer (2011).Búsqueda colaborativa en la ficha clínica electrónica

[3] Amir Gandomi & Murtaza Haider (2014). Beyond the hype-Big data concepts, methods, and analytics

[4] Fan J., Han F. & Liu H. (2014) Challenges of Big data analysis.

[5] Knowbot [Internet] (2015) https://en.wikipedia.org/wiki/Knowbot

[6] IBM (2012) ¿Qué es Big Data?

[7] Purcell, Bernice (2013). The emergence of Big Data technology and Analytics

[8] Pérez IJ, Cabrerizo FJ, Alonso S, Herrera-Viedma E (2014) A new consensus model for group decision making problems with non-homogeneous experts.

[9] Xu ZS, Da QL (2004) An overview of operators for aggregating information.

[10] Cabrerizo FJ, Herrera-Viedma E, Pedrycz W (2013) A method based on PSO and granular computing of linguistic information to solve group decision making problems defined in heterogeneous contexts.

[11]Herrera F, Alonso S, Chiclana F, Herrera-Viedma E (2009) Computing with words in decision making: foundations, trends and prospects.

[12] HocevarKP, Flanagin AJ,MetzgerMJ(2014) Social media self-efficacy and information evaluation online.

[13] Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of Social Media.

Anexos Técnicos

1. Descripción de entidades y atributos

ENT –01	Información
Descripción	Llamaremos INFORMACIÓN a toda porción de texto sin procesar por el sistema que se ha leído como una unidad de algún de las fuentes consideradas. Puede ser de varios tipos en función de la fuente de donde se ha extraído.
Atributos	ATR – 01: Informacion_id ATR – 02: Fecha ATR – 03: Hora
Comentarios	Ninguno
ATR – 01	INFORMACIÓN :: Informacion_id
Descripción	Identificador único para cada porción de información
Tipo	Big Integer
Comentarios	Clave primaria, auto-incremental
ATR – 02	INFORMACIÓN :: Fecha
Descripción	Fecha de la publicación de la información según la fuente
Tipo	Date
Comentarios	No nulo
ATR – 03	INFORMACIÓN :: Hora
Descripción	Indica la hora de publicación de la información según el medio de donde se ha leído
Tipo	Time
Comentarios	Optativo

ENT - 02	Noticia
Descripción	Una noticia es un texto con la información leída de la página web de un medio de comunicación.
Supertipos	INFORMACION (ENT – 01)
Atributos	ATR – 04: Titulo ATR – 05: Cabecera ATR – 06: URL ATR – 07: Cuerpo ATR – 08: Autor
Comentarios	Hereda la clave primaria de ENT – 01
ATR – 04	NOTICIA :: Titulo
Descripción	Almacena el Titulo de la Noticia
Tipo	Varchar
Comentarios	No nulo
ATR – 05	NOTICIA :: Cabecera
Descripción	Almacena el Cabecera de la Noticia
Tipo	Varchar
Comentarios	No nulo
ATR – 06	NOTICIA :: URL
Descripción	Almacena el URL de la Noticia
Tipo	Varchar

Comentarios	No nulo
ATR – 07	NOTICIA :: Cuerpo
Descripción	Almacena el Cuerpo de la Noticia
Tipo	Varchar
Comentarios	Optativo
ATR – 08	NOTICIA :: Autor
Descripción	Almacena el Autor de la Noticia
Tipo	Varchar
Comentarios	Optativo

ENT - 03	Menéame
Descripción	Llamaremos Menéame a los datos extras que ofrece la página web y que no pueden ser almacenados en ENT – 02
Supertipos	NOTICIA (ENT – 02)
Atributos	ATR – 09: Usuario ATR – 10: Clics ATR – 11: Meneos ATR – 12: NComentarios
Comentarios	Hereda la clave primaria de ENT – 01
ATR – 09	MENEAME :: Usuario
Descripción	Almacena el nombre del Usuario
Tipo	Varchar
Comentarios	Optativo
ATR – 10	MENEAME :: Clics
Descripción	Almacena el numero de Clics hecho por los usuarios
Tipo	Integer
Comentarios	Optativo
ATR – 11	MENEAME :: Meneos
Descripción	Almacena el número de Meneos que tiene la noticia. Dar meneos es la forma de votar una noticia en la página menéame.
Tipo	Integer
Comentarios	Optativo
ATR – 12	MENEAME ::NComentarios
Descripción	Almacena el Cuerpo de la Noticia
Tipo	Integer
Comentarios	Optativo

ENT –04	Twitter
Descripción	En la entidad Twitter se podrá encontrar cada unidad de información leída de Twitter.
Supertipos	INFORMACION (ENT – 01)
Atributos	ATR – 13: Tweet_id ATR – 14: Texto ATR – 15: Retweets ATR – 16: EsRetweet ATR – 17: EnRespuestaA

Comentarios	Hereda la clave primaria de ENT – 01
ATR – 13	TWITTER ::Tweet_id
Descripción	Identificador único de cada tweet.
Tipo	Big Integer
Comentarios	Clave primaria, auto-incremental
ATR – 14	TWITTER :: Texto
Descripción	Texto que contiene el tweet
Tipo	Varchar
Comentarios	Obligatorio
ATR – 15	TWITTER ::Retweets
Descripción	Numero de retweets que tiene el tweet
Tipo	Integer
Comentarios	Obligatorio
ATR – 16	TWITTER ::EsRetweet
Descripción	Indica si el tweet es un retweet
Tipo	Boolean
Comentarios	Obligatorio.
ATR – 17	TWITTER ::EnRespuestaA
Descripción	Indica el tweet_id al que va dirigida este tweet
Tipo	Big Integer
Comentarios	Optativo

ENT - 05	Tema
Descripción	Llamaremos Tema al 1º post de cada tema de discusión que guardemos
Supertipos	INFORMACION (ENT – 01)
Atributos	ATR – 18: Titulo ATR – 19: Texto ATR – 20: URL_Inicial ATR – 21: URL_Final
Comentarios	Hereda la clave primaria de ENT – 01
ATR – 18	TEMA :: Titulo
Descripción	Almacena el Titulo del Tema
Tipo	Varchar
Comentarios	Obligatorio
ATR – 19	TEMA :: Texto
Descripción	Almacena el primer post del Tema
Tipo	Varchar
Comentarios	Obligatorio
ATR – 20	TEMA :: URL_Inicial
Descripción	Almacena la url de la primera página del tema
Tipo	Varchar
Comentarios	Obligatoria
ATR – 21	TEMA :: URL_Final
Descripción	Almacena la url de la última página del tema leída por el sistema
Tipo	Varchar

Comentarios	Obligatoria

ENT - 06	Post
Descripción	Post almacenara todas las respuestas a los temas de discusión
Supertipos	INFORMACION (ENT – 01)
Atributos	ATR – 22: Texto ATR – 23: Tema_id ATR – 24: Post_ordem
Comentarios	Hereda la clave primaria de ENT – 01
ATR – 22	POST :: Texto
Descripción	Almacena el texto de los Post
Tipo	Varchar
Comentarios	Obligatorio
ATR – 23	POST :: Tema_id
Descripción	Indica a que tema pertene el post
Tipo	Big Integer
Comentarios	Obligatorio
ATR – 24	POST :: Post_orden
Descripción	Indica la posición del post en el Tema
Tipo	Varchar
Comentarios	Obligatorio

ENT - 07	Localizacion
Descripción	Indica la procedencia
Atributos	ATR – 25: Localizacion_id ATR – 26: Nombre ATR – 27: Pais_id
ATR – 25	LOCALIZACION :: localización_id
Descripción	Identificador único de cada localización
Tipo	Integer
Comentarios	Clave principal, auto-incremental
ATR – 26	LOCALIZACION :: Nombre
Descripción	Nombre de la localización
Tipo	Varchar
Comentarios	Obligatoria
ATR – 27	LOCALIZACION :: Pais_id
Descripción	Indica si la localización pertenece a otra localización
Tipo	Integer
Comentarios	Ninguno

ENT –08	Idioma
Descripción	Indica un idioma
Atributos	ATR –28: Idioma_id ATR –29:Codigo ATR –30:Nombre

ATR – 28	IDIOMA :: Idioma_id
Descripción	Identificador único de cada idioma
Tipo	Integer
Comentarios	Obligatorio, auto-incremental
ATR – 29	IDIOMA :: Codigo
Descripción	Abreviación del nombre
Tipo	Varchar
Comentarios	Obligatorio
ATR – 30	IDIOMA :: Nombre
Descripción	Nombre del idioma
Tipo	Varchar
Comentarios	Obligatorio

ENT - 09	Fuente
Descripción	En esta entidad se almacenaran los nombres de los sitios de donde se obtiene la información
Atributos	ATR – 31: Fuente_id ATR – 32: Descripcion ATR – 33: Entidad_id
ATR – 31	FUENTE :: Fuente_id
Descripción	Identificador único de cada fuente
Tipo	Integer
Comentarios	Obligatorio, auto-incremental
ATR – 32	FUENTE :: Descripcion
Descripción	Nombre de la fuente
Tipo	Varchar
Comentarios	Obligatorio
ATR – 33	FUENTE :: Entidad_id
Descripción	Aquí se almacena si la fuente es también alguna entidad
Tipo	Integer
Comentarios	Optativa

ENT - 10	RSS
Descripción	Aquí se almacenaran las RSS de la prensa digital
Atributos	ATR – 34: URL
ATR – 34	RSS :: URL
Descripción	Aquí se almacenan las url que nos llevan a las RSS de la prensa digital
Tipo	Varchar
Comentarios	Clave primaria

ENT - 11	TipoFuente
Descripción	Aquí se almacenaran los diferentes tipos de fuentes de donde podemos leer
Atributos	ATR – 35: TipoFuente ATR – 36: Descripcion

ATR – 35	TIPOFUENTE :: TipoFuente
Descripción	Identificador único de cada TipoFuente
Tipo	Integer
Comentarios	Obligatorio, auto-incremental
ATR – 36	TIPOFUENTE :: Descripcion
Descripción	Especificación de de la fuente indicando si proviene de un presna, twitter, foros,…
Tipo	Varchar
Comentarios	Obligatorio

ENT - 12	Sinonimo
Descripción	Aquí almacenaremos los sinónimos de las entidades
Atributos	ATR – 37: Sinonimo_id ATR – 38: Sinonimo_str ATR – 39: Frecuente ATR – 40: SinonimoParaFuente
ATR – 37	SINONIMO :: Sinonimo_id
Descripción	Identificador único de cada TipoFuente
Tipo	Integer
Comentarios	Obligatorio, auto-incremental
ATR – 38	SINONIMO :: Sinonimo_str
Descripción	Aquí almacenaremos el nombre del sinónimo
Tipo	Varchar
Comentarios	Obligatorio
ATR – 39	SINONIMO :: Frecuente
Descripción	Indicaremos si el sinónimo es frecuente o no
Tipo	Boolean
Comentarios	Obligatorio
ATR – 40	SINONIMO :: SinonimoParaFuente
Descripción	Aquí almacenaremos si el sinónimo es exclusivo para alguna fuente
Tipo	Integer
Comentarios	Optativo

ENT - 13	Anti-Sinonimo
Descripción	Aquí almacenaremos los Anti-Sinonimos de los sinónimos
Atributos	ATR – 41:Antisinonimo_str
ATR – 41	ANTI-SINONIMO :: Antisinonimo_str
Descripción	Nombre del Anti-Sinonimo
Tipo	Varchar
Comentarios	Clave Primaria

ENT - 14	Entidad
Descripción	Bajo el concepto de ENTIDAD almacenaremos cualquier empresa, persona u organismo, del que bien nos interesa almacenar información, o bien publica información de una

	entidad que sí nos interesa.
Atributos	ATR – 42: Entidad_id ATR – 43: Identificador ATR – 44: Objetivo
ATR – 42	ENTIDAD :: Entidad_id
Descripción	Identificador para cada una de las entidades
Tipo	Integer
Comentarios	Obligatorio, auto-incremental
ATR – 43	ENTIDAD :: Identificador
Descripción	Nombre de la entidad
Tipo	Varchar
Comentarios	Obligatorio, único.
ATR – 44	ENTIDAD :: Objetivo
Descripción	Booleano que nos indica si la empresa es
Tipo	Boolean
Comentarios	Obligatorio, iniciado en false.

ENT - 15	Empresa
Descripción	Esta entidad indica que Entidades son una empresa
Atributos	Ninguno
Supertipo	Entidad (ENT – 14)

ENT - 16	Persona
Descripción	Esta entidad indica que Entidades son una persona
Atributos	ATR – 45: Nombre ATR – 46: Apellidos ATR – 47: Busqueda_str ATR – 48: Busqueda_patron ATR – 49: Cargo
Supertipo	Entidad (ENT – 14)
ATR – 45	PERSONA :: Nombre
Descripción	Aquí se guardara el nombre de la persona
Tipo	Varchar
Comentarios	Obligatorio
ATR – 46	PERSONA :: Apellidos
Descripción	Aquí se guardara el Apellido(s) de la persona
Tipo	Varchar
Comentarios	Obligatorio
ATR – 47	PERSONA :: Busqueda_str
Descripción	Aquí se guardara el patrón de búsqueda para la persona
Tipo	Varchar
Comentarios	Obligatorio
ATR – 48	PERSONA :: Busqueda_patron
Descripción	Aquí se guardara el patrón de búsqueda para la persona específico para twitter.
Tipo	Varchar
Comentarios	Obligatorio

ATR – 49	PERSONA :: Cargo
Descripción	Aquí se guardara el Cargo de la persona en la empresa
Tipo	Varchar
Comentarios	Optativo

ENT - 17	Patrones
Descripción	Aquí almacenaremos los patrones de búsqueda.
Atributos	ATR – 50: Patron_id ATR – 51:Patron
ATR – 50	PATRONES :: Patron_id
Descripción	Identificador para cada uno de los patrones
Tipo	Integer
Comentarios	Clave primaria, auto-incremental
ATR – 51	PATRONES :: Patron
Descripción	Expresion Regular que forma el patrón
Tipo	Varchar
Comentarios	Obligatorio

ENT - 18	Patron_Posicion
Descripción	Aquí almacenamos en que posición se aplicara el patrón
Atributos	ATR – 52: Patron_posicion_id ATR – 53: Posicion
ATR – 52	PATRON_POSICION :: Patron_posicion_id
Descripción	Identificador para cada una de las posiciones
Tipo	Integer
Comentarios	Clave Primaria, auto-incremental.
ATR – 53	PATRON_POSICION :: Posicion
Descripción	Este campo indicara en qué posición se aplicara el patron
Tipo	Integer
Comentarios	Obligatorio

ENT - 19	TipoPatron
Descripción	Esta entidad indicara para que se usa el patrón, si para una notica o foro.
Atributos	ATR – 54: Tipo_id ATR – 55: Descripcion
ATR – 54	TIPOPATRON :: Tipo_id
Descripción	Identificador para cada uno de los TipoPatron
Tipo	Integer
Comentarios	Clave Primaria, auto-incremental
ATR – 55	TIPOPATRON :: Descripcion
Descripción	Nombre del TipoPatron
Tipo	Varchar
Comentarios	Obligatorio

ENT - 20	Usuario
Descripción	En esta Entidad se almacenaran los usuarios del sistema web
Atributos	ATR – 56: Usuario_id ATR – 57: User ATR – 58: Pass ATR – 59: Activo ATR – 60: Entidad_id
ATR – 56	USUARIO :: Usuario_id
Descripción	Identificador para cada uno de los Usuarios
Tipo	Integer
Comentarios	Clave Primaria, auto-incremental
ATR – 57	USUARIO ::User
Descripción	Nombre del usuario
Tipo	Varchar
Comentarios	Obligatoro, único
ATR – 58	USUARIO :: Pass
Descripción	Password del usuario
Tipo	Varchar
Comentarios	Obligatoria
ATR – 59	USUARIO :: Activo
Descripción	Estado en el que se encuentra la cuenta
Tipo	Boolean
Comentarios	Obligatorio, se inicializa en False
ATR – 60	USUARIO :: Entidad
Descripción	Empresa a la que está asociado el usuario
Tipo	Big Integer
Comentarios	Optativo

ENT - 21	Perfil
Descripción	Aquí se almacenara los perfiles de usuarios
Atributos	ATR – 61: Perfil_id ATR – 62: Descripcion
ATR – 61	PERFIL :: Perfil_id
Descripción	Identificador para cada uno de los Perfiles
Tipo	Integer
Comentarios	Clave Primaria, auto-incremental
ATR – 62	PERFIL :: Descripcion
Descripción	Nombre del perfil
Tipo	Varchar
Comentarios	Obligatorio

2. Descripción de relaciones entre entidades

REL – 01	**Origen**	ENT–01	**Mult**	**Destino**	ENT–12	**Mult**
Definida sobre	Informacion		1…*	Sinonimo		1…*
Calificador	Tiene			Aparece		
Descripción	Una información tiene uno o más sinónimos Un sinónimo aparece en una o más informaciones					
Optatividad	No			No		

REL – 02	**Origen**	ENT–01	**Mult**	**Destino**	ENT–07	**Mult**
Definida sobre	Informacion		1	Localizacion		*
Calificador	Tiene			Tiene		
Descripción	Una información tiene una localización Una localización tiene una o más información					
Optatividad	No			Si		

REL – 03	**Origen**	ENT–01	**Mult**	**Destino**	ENT–09	**Mult**
Definida sobre	Informacion		1	Fuente		*
Calificador	Pertenece			Tiene		
Descripción	Una información pertenece a una fuente Una fuente tiene cero o más informaciones					
Optatividad	No			Si		

REL – 04	**Origen**	ENT–01	**Mult**	**Destino**	ENT–08	**Mult**
Definida sobre	Informacion		1	Idioma		*
Calificador	Esta escrito			Tiene		
Descripción	Una Informacionestá escrito en un idioma Un idioma tiene cero omásinformaciones					
Optatividad	No			Si		

REL – 05	**Origen**	ENT–09	**Mult**	**Destino**	ENT–10	**Mult**
Definida sobre	Fuente		*	RSS		1
Calificador	Tiene			Pertenece		

Descripción	Una fuente puede tener cero o más RSS Un RSS solo puede tener una fuente	
Optatividad	Si	No

REL – 06	**Origen**	ENT–09	**Mult**	**Destino**	ENT–11	**Mult**
Definida sobre	Fuente		1	TipoFuente		*
Calificador	Tiene			Define		
Descripción	Una fuente tiene un tipo fuente Un tipofuente puede definir a cero o más fuentes.					
Optatividad	No			Si		

REL – 07	**Origen**	ENT–05	**Mult**	**Destino**	ENT–06	**Mult**
Definida sobre	Tema		*	Post		1
Calificador	Tiene			Pertenece		
Descripción	Un tema puede tener cero o más post Un post pertenece a un tema					
Optatividad	Si			No		

REL – 08	**Origen**	ENT–12	**Mult**	**Destino**	ENT–13	**Mult**
Definida sobre	Sinonimo		*	Anti-Sinonimo		1
Calificador	Tiene			Hace referencia		
Descripción	Un Sinonimo puede tener cero o mas Anti-Sinonimo Un Anti-Sinonimo hace referencia a un Sinonimo					
Optatividad	Si			No		

REL – 09	**Origen**	ENT–14	**Mult**	**Destino**	ENT–12	**Mult**
Definida sobre	Entidad		*	Sinonimo		1
Calificador	Tiene			Hace referencia		
Descripción	Una Entidad puede tener cero o masSinonimos Un Sinonimo hace referencia a una sola Entidad					
Optatividad	No			No		

REL – 10	Origen	ENT–17	Mult	Destino	ENT–18	Mult
Definida sobre	Patrones		1…*	Patron_Posicion		1…*
Calificador	Es posicionado			Posiciona		
Descripción	Un patrón posicionado por 1 o más Patron_Posicion Un Patron_Posicion posiciona uno o más Patrones					
Optatividad	No			No		

REL – 11	Origen	ENT–18	Mult	Destino	ENT–19	Mult
Definida sobre	Patron_Posicion		1	TipoPatron		*
Calificador	Es catalogado			cataloga		
Descripción	Un Patron_Posicion es catalogado por un TipoPatron Un TipoPatron cataloga cero o masPatron_Posicion					
Optatividad	Si			No		

REL – 12	Origen	ENT–20	Mult	Destino	ENT–21	Mult
Definida sobre	Usuario		1	Perfil		*
Calificador	Tiene			Pertenece		
Descripción	Un usuario tiene un perfil Un perfil pertenece a cero o mas usuarios					
Optatividad	No			Si		