UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MÁSTER UNIVERSITARIO EN INTELIGENCIA ARTIFICIAL, RECONOCIMIENTO DE FORMAS E IMAGEN DIGITAL

TESINA DE MÁSTER

SISTEMA DE RECUPERACIÓN DE INFORMACIÓN COMO MOTOR DE LA ESCUCHA ACTIVA

Autor:
Víctor Manuel Dart Andreu

Directores:
Francisco M. Rangel Pardo
Autoritas Consulting (*)

Dr. Paolo Rosso
Universitat Politècnica de València

Valencia, 2012

(*) Director de Tecnología en Autoritas Consulting

RESUMEN

La vertiginosa evolución de las tecnologías de información, la aparición de Internet y las evoluciones sufridas en la Web, han provocado el surgimiento de una serie de aplicaciones colaborativas donde los usuarios no solo consumen información.

Debido a las características propias de este tipo de herramientas Web, los usuarios pueden crear, eliminar o modificar nuevos contenidos, aportando nueva información a la red. Así pues, Internet se ha convertido en la primera fuente de información y comunicación para los usuarios. No obstante, debido al crecimiento exponencial de Internet, a la innumerable cantidad de usuarios que lo componen y a las nuevas posibilidades de colaboración, se están generando grandes cantidades de información desestructuradas que dificultan la búsqueda y recuperación de información relevante para un usuario.

La complejidad de esto, hace que cada vez resulte más complicado buscar, utilizar y compartir información útil. El mundo empresarial no es ajeno a este nuevo paradigma. Las nuevas tecnologías han permitido que Internet se convierta en una red llena de posibilidades, donde la comunicación y el intercambio de contenidos han revelado un nuevo modelo de estrategia de marketing. Las empresas tienen mayores medios donde hacer publicidad, aumentando así las posibilidades de encontrar a usuarios interesados en sus servicios. Además, las nuevas vías de comunicación permiten a estos expresar sus opiniones e inquietudes, revelando así, información que puede ser de gran interés para las empresas, y que han de saber aprovechar las innumerables ventajas que ofrece este nuevo paradigma de comunicación.

Es por ello que se plantea la exigencia de un sistema de recuperación de información en Internet como módulo imprescindible de una aplicación de escucha activa que ayude a las empresas en el análisis de la información para la toma adecuada de sus decisiones.

Así pues, mediante el empleo de técnicas de recuperación de información de las plataformas Web 2.0, se ha desarrollado un sistema que mediante la utilización de las interfaces de programación disponibles en algunas de las herramientas 2.0 se pueda obtener la mayora cantidad de información relevante para una necesidad de información por parte del usuario. El sistema atiende las peticiones propuestas para la escucha activa por el usuario y activa la recuperación de información centrada en la utilización de diferentes interfaces y la combinación de *crawlers* específicos que ayuden a la mejor comprensión del contenido. Para ello, se ha divido el sistema en tres módulos que engloban a las fuentes de información con más relevancia para el proyecto. En el primero, encontramos las fuentes de tipo texto tales como blog, prensa digital, foros y otras Web. En el segundo, engloba las fuentes de tipo redes sociales o microblogging, tales como Twitter o Facebook, mientras que el último módulo tiene las fuentes multimedia como YouTube y Flickr.

El sistema de recuperación obtiene información de las distintas fuentes con la finalidad de ofrecer al usuario una información variada y de relevancia para su posterior análisis en los distintos módulos de la escucha activa.

El trabajo realizado en esta tesina se ha desarrollado como módulo indispensable de un proyecto de escucha activa para la empresa Autoritas Consulting, bajo la supervisión del director de tecnología de esta.

AGRADECIMIENTOS

A Barrio Sésamo por su influencia en el cambio socio-cultural de esa generación,
A Coco por enfocarme en la teoría de la entropía,
A mis padres por darme la oportunidad,
A mi novia por su paciencia,
A Kiko que más que ser mi tutor en esta tesina, es un amigo y compañero filosófico de teorías *frikys*,
A Paolo Rosso por su tiempo y dedicación,
A mis compañeros de trabajo por su insistencia y apoyo,
A todo el mundo que ha creído en mí y en que podía hacerlo.

Los trabajos de investigación que han dado lugar a esta tesina han sido parcialmente financiados por los proyectos del ministerio ITC/464/2008 y TSI-020100-2011-56 concedidos a Autoritas Consulting

ÍNDICE

ÍNDICE DE FIGURAS

1. INTRODUCCIÓN

1.1 Motivación

Los avances tecnológicos de los últimos sesenta años han hecho posible la globalización social o mundialización, donde la información es uno de los recursos más importantes de la cultura y la sociedad actual. Esto ha provocado la aparición de un nuevo término para describir el modelo de comunicación actual, la sociedad red [B.1].

Nos hallamos sumergidos en la edad de oro de la información, cada vez tenemos más cantidad de información disponible y mayores medios para acceder a ella. El proceso de digitalización de documentación así como el desarrollo de las nuevas tecnologías de la información son claros ejemplos de la revolución de la información, cuyos sistemas tienden a superarse y mejorarse día a día.

Los medios de comunicación actuales suponen nuevos modelos de transferencia, admisión y procedimiento de la información, por lo que ofrecen nuevos espacios de acción, nuevas ocasiones de trabajo y nuevos mercados por explorar. Además, la incorporación de las nuevas tecnologías ha hecho que el trabajo sea más eficaz y contemple estándares que aseguren el funcionamiento óptimo y la calidad de adaptación a las constantes evoluciones.

La expansión de estos medios y concretamente las redes informáticas, han permitido las relaciones entre entidades de distintos puntos del mundo, aumentando así el intercambio de tecnologías en su aspiración de incrementar su competitividad. Este conocimiento tecnológico está estrechamente ligado con el progreso económico y productivo, ya que definen en cada momento la estructura y el dinamismo del sistema, por ello se puede afirmar que la tecnología se ha convertido en un elemento indispensable de productividad y herramienta de competitividad para cualquier entidad.

A partir de esta sociedad de información nacen conceptos como la gestión del conocimiento y el comercio digital. Actualmente Internet es el medio de información más grande conocido, siendo una de las principales fuentes de generación y transmisión de datos. No obstante, debido a este crecimiento, su problema principal es el descontrol de la información a la que los usuarios pueden acceder, provocando dificultades para encontrar de manera simple y eficiente información de calidad.

Centrándonos en el modelo empresarial, esta revolución de información obliga a cambiar el esquema tradicional de los procesos de gestión y productivos, obligando a las empresas a ser más competitivas e innovadoras tecnológicamente hablando.

La gran cantidad de posibilidades que ofrece este medio, entre ellas las herramientas colaborativas como redes sociales, *blogs*, *wikis*...han creado un nuevo tipo de Web, la Web social [B.2], donde los usuarios no solo consumen información, sino que también son productores de esta.

Toda esta comunicación está provocando un crecimiento exponencial de información no estructurada en Internet, ofreciendo una oportunidad a las empresas de captarla y establecer una actitud de escucha activa frente a los usuarios, atendiéndolos individualmente, dándoles respuestas y entendiendo sus inquietudes. Ante esta situación la empresa ha de saber actuar respondiendo de forma efectiva y eficaz a los retos que presenta la globalización y las competencias actuales, aprovechando las oportunidades de la sociedad red que se le brindan y así aventurar una mayor innovación en el nuevo contexto.

Existe una estrecha relación entre la revolución de información, la sociedad de información e Internet, debido principalmente a que este último permite un flujo de dimensiones indeterminadas de información. Pero, ¿Quién se encarga de generar y consumir la información disponible en este medio?

En respuesta a esta pregunta, aparece un nuevo concepto: los prosumidores [B.3] (en inglés *prosumer*). A diferencia de otros medios de comunicación y gracias a la libertad que brindan los nuevos medios digitales, todos los usuarios pueden ofrecer información, elaborar contenidos y ser consumidores de los mismos, convirtiéndose en generadores de contenidos por excelencia.

Hasta esta nueva sociedad, las únicas transmisoras de información se centraban en ámbito familiar y en los medios de comunicación masivos. Por entonces, la propagación de esta información cumplía con dos objetivos: la transmisión de la información y la creación de una opinión común. Sin embargo, con la llegada de las nuevas tecnologías se ha aumentado el crecimiento de los canales de comunicación y por ende, el incremento del conocimiento.

La sociedad de información está en continua evolución. La comunicación y la información se han convertido en los bienes más preciados, debido a que el eje central de dicha evolución es la globalización, la cual erosiona las fronteras y modifica los ejes en las tomas de decisión.

Esta nueva realidad está concretando un nuevo modelo organizativo, así como un nuevo talante estratégico y social, en el que es necesario realizar sistemas que ayuden a las empresas y profesionales a tratar la gran cantidad de información, mediante la escucha activa para tener constancia de qué se dice, cuándo se dice, cómo se dice y dónde se dice. En conclusión, la información ha pasado de ser una herramienta de producción de bienes, ha convertirse en el principal de los bienes de una empresa.

1.2 Gestión del conocimiento

De las nociones de la sociedad de información nace la noción de gestión de conocimiento, en ella la información se registra como un factor de éxito en el mundo empresarial, por lo que las empresas deben de centrar sus intereses en desarrollar una estrategia competitiva basada en instrucciones e instrumentos que permitan recoger, examinar y utilizar la información para lograr un valor añadido de ella.

Las bases de la gestión del conocimiento pueden ser vistos desde dos puntos globales [B.4]:

- La información es la base del conocimiento, por lo que debe de obtenerse, analizarse y explotarse para alcanzar los objetivos de la empresa.

- La gestión de esta información debe de ser un proceso cotidiano de las actividades diarias de la empresa.

Basándose en estos fundamentos, la gestión de conocimiento puede ser definida como el proceso por el cual se ha de encontrar, obtener, analizar y presentar la información para ser transformada en conocimiento, con el fin de tener una mayor idea del entorno y poder dar una respuesta desde la propia experiencia de la empresa.

Para una buena administración de la información se ha de saber escuchar. En el proceso de comunicación, uno de los elementos más difícil e importante, es la escucha y la retroalimentación. Sin ellos, no hay comunicación. Actualmente, la falta de comunicación que se está sufriendo es principalmente por la falta de no saber escuchar a los demás. Se pierde más tiempo en difundir y seguir las propias emisiones, que en analizar las respuestas a esta información.

La escucha activa, no es un concepto nuevo y una de sus definiciones fuera del ámbito digital puede ser el proceso por el cual una persona escucha y entiende la comunicación desde el punto de vista del que habla, percibiendo los sentimientos, ideas o pensamientos que surgen de lo que se está expresando. Para alcanzar a entender a la persona que se está comunicando es necesario algo de empatía, o sea, se ha de saber poner en el lugar de la otra persona. No hay que confundir el término oír con escuchar: mientras que el oír es

únicamente percibir vibraciones de sonido en el oído, el escuchar, es razonar, dar sentido y entender lo que se oye.

Si trasladamos esta definición al mundo digital y desde un punto de vista empresarial, la escucha activa se divide en una serie de conceptos claves, en función a la información que sea relevante para una empresa: monitorizar, capturar, filtrar, analizar, catalogar, pensar, valorar y actuar, son los conceptos claves de este nuevo término. La escucha activa representa el escuchar y entender la comunicación desde el lugar del que habla, convirtiéndose actualmente en un proceso fundamental en las empresas para realizar estudios de marketing.

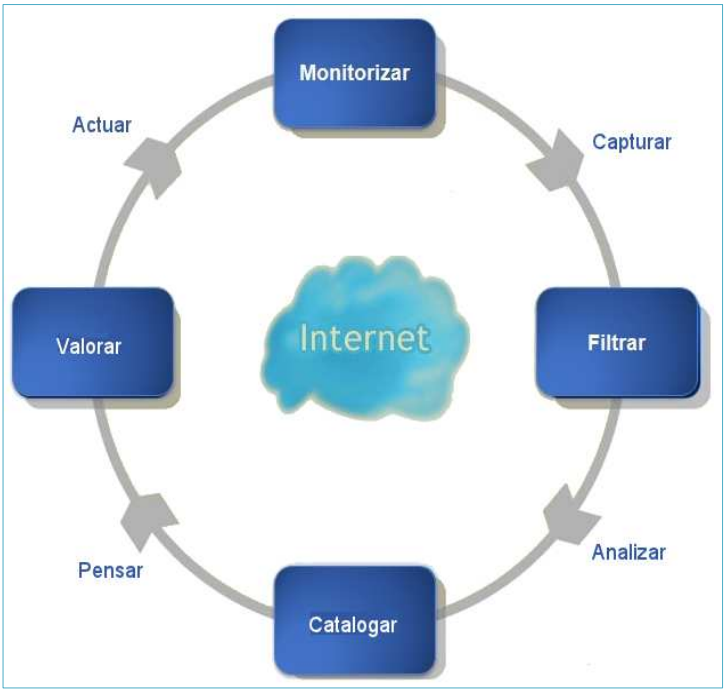

Figura 1. Estructura de la escucha activa en Internet

Existen muchas empresas que recogen mensajes de los usuarios, pero, ¿Quién escucha y responde a las necesidades de estos? Estamos en la sociedad del consumo con infinidad de servicios y productos, la noción de fidelidad en la marca se ha transformado y la exigencia del consumidor ha aumentado notablemente. Por ello, nos encontramos en una situación en que se debe cuidar y escuchar a los clientes, comprendiendo el mensaje que nos transmite, interpretando los gestos y entendiendo más allá de sus palabras.

Numerosos son los medios en que los usuarios expresan sus opiniones, siendo las redes sociales el principal canal de comunicación entre el consumidor y las empresas. Desde los perfiles sociales, los usuarios siguen a las organizaciones con las que se sienten identificados y tienen una afinidad común expresando sus opiniones frente a estas, ya sea premiándolas cuando realizan una buena gestión, manifestando sus desacuerdos cuando hagan una mala acción o simplemente para consultar información.

En este punto es cuando las empresas entran en disputa y en entredicho. La reacción de la empresa y su forma de actuar harán que el usuario cambie su postura frente a esta.

El hecho es que actualmente muchas de las organizaciones no disponen de un sistema de escucha activa que les permita conocer la opinión de los usuarios o no están presentes en los diferentes canales de información.

1.3 Objetivos

Las personas no podemos captar toda la información que se encuentra en la red, por lo que necesitamos herramientas que nos permitan obtener y tratar está información.

El objetivo principal del proyecto es la creación de una herramienta totalmente modular de recuperación de información para la escucha activa en Internet. Que mediante la integración de sus módulos independientes, permita la captación de información notable para una empresa, pudiéndole servir como herramienta de apoyo en la evaluación de su impacto en la Web y los *social media* [B.5].

Mediante el empleo de técnicas de recuperación de información de las plataformas Web 2.0, el sistema dispondrá de un módulo de consulta simple y otro avanzado, donde el usuario satisfará sus necesidades de información, introduciendo palabras claves así como filtros que mejoraran los resultados obtenidos. El sistema se encargará de interpretar las consultas introducidas por el usuario y obtener los mejores resultados respecto a la necesidad de información. El flujo de información será continuo siempre y cuando el usuario no descarte la consulta de información.

Adicionalmente, se requerirá que el sistema sea efectivo en las búsquedas, y la eficiencia de este será evaluada en base al rendimiento global, su agilidad y la rapidez con que devuelve los resultados. Además, se tendrá muy en cuenta el sistema de interfaz presentado al usuario, ya que es la parte que ven estos y el éxito o no de una herramienta viene ligado muchas veces por este punto.

1.4 Estructura de la memoria

El proyecto está compuesto por la siguiente estructura, la cual está dividida en cinco capítulos.

En el capítulo uno hemos visto una introducción de los avances tecnológicos como motores del cambio social y de las necesidades de las empresas. Estos cambios tecnológicos y sociales, están provocando un crecimiento de la información desestructurada en Internet, acrecentando, la importancia de las herramientas de escucha activa debido a la necesidad de información por parte de las empresas para la toma de decisiones.

En el capítulo dos se hace un recorrido por el concepto de la Web 2.0. Mediante una breve historia, se intenta introducir al usuario en los orígenes de la Web (en inglés, *World Wide Web*) hasta llegar al nacimiento de la actual Web, la 2.0. A continuación, se exponen los usos y aplicaciones más característicos de esta plataforma, destacando entre ellos: las redes sociales, los *blogs*, los foros y las herramientas de intercambio de contenidos.

En el capítulo tres se hace una revisión al estado del arte donde se enmarca el contexto sobre el que se desarrolló el proyecto. Además se realiza una visión general de los sistemas de recuperación de información, viendo cuales son sus modelos, que problemas surgen a la hora de cambiar el entorno de recuperación en estos sistemas y cuales son las principales fuentes de información actualmente. También, se analizan algunas herramientas de escucha activa donde veremos algunas de sus características y la importancia de este tipo de sistema para las empresas.

En el capítulo cuatro se hace una visión general de los puntos más importantes de la aplicación de escucha activa Cosmos.

En el capítulo cinco se hace un recorrido por la tecnología utilizada. Debido a la variedad de estas, se hace conveniente hacer este capitulo donde se da una visión general de las características de cada una de ellas.

En el capítulo seis se describe el módulo de recuperación de información, donde se define toda la implementación y la descripción de los diferentes sistemas utilizados.

Por último, en el capítulo siete se presentará las conclusiones a las que se ha llegado al implementar el sistema.

2. WEB 2.0

2.1 El origen

Aproximadamente quince años después de que el filósofo y sociólogo Ted Nelson [B.6] pensara en la idea de que todas las máquinas del mundo pudieran publicar información en hipertexto, un primerizo ingeniero Tim Berners-Lee [B.7], estudiaba como hacer real este proyecto. Para ello, estuvo investigando en un sistema de difusión de información descentralizada apoyado en el hipertexto.

Durante los años 1988 y 1989 se realizaron diferentes experimentos para poder comunicar los centros de física nuclear de todo el mundo. A raíz de esto, se intentaron múltiples protocolos de comunicación, definiéndose el estándar para protocolos de conexiones TCP-IP (del inglés, *Transmission Control Protocol – IP*). Es entonces, cuando Tim Berners-Lee, plantea la utilización de un sistema de intercambio de información fundado en el hipertexto.

Alrededor del año 1990, Tim Berners-Lee y Robert Cailliau [B.8] crearon la Web, vocablo del inglés que significa red, telaraña o malla. En sus comienzos se utilizaban páginas estáticas programadas en HTML (del inglés, *Hyper Text Markup Language*) las cuales carecían principalmente de su falta de actualización. Este sistema permitía el acceso a los distintos recursos que ofrecía Internet a través de una interfaz común basada en el hipertexto, donde su estructura básica sobre documentos y enlaces, se diseñó para la lectura por humano y no para que el contenido fuese procesado de forma automática.

La Web 1.0 únicamente es de lectura, el usuario únicamente puede obtener información mediante la interacción sobre esta, estando totalmente limitado a los contenidos que el creador de la página publique, sin poder participar activamente en la interacción con la información. El éxito de estas páginas dependía fuertemente de su dinamismo, por lo que los sistemas de gestión de contenidos ofrecían paginas HTML erigidas al momento desde una base de datos actualizada. La Web 1.0 evolucionó a lo que se llamó páginas Web 1.5, donde se introdujeron nuevas tecnologías como el ASP (del inglés *Active Server Pages*) o CSS (del inglés *Cascading Style Sheets*), cuyos principales factores consistían en la estética visual y la cantidad de visitas (*hits*) conseguidas.

La principal característica en la Web 1.5 es la introducción del dinamismo en las páginas Web ofreciendo la posibilidad de mantener la información actualiza por los usuarios a través de editores de contenidos. Es entonces cuando comienzan a aparecer los primeros *Weblogs* [B.9] colaborativos.

Las dos últimas décadas del siglo pasado asistimos al primer ciclo de auge y recesión de los modelos de negocio desarrollados sobre Internet y la Web. El comienzo de este período se inició con la salida a bolsa de la empresa *Netscape Communications Corporation* [B.10] en 1995. Desde aquél momento, se produjo un incremento exponencial de nuevos negocios desarrollados en la Web, originando transacciones de muchísimo dinero. A esta etapa se la conoce como la *burbuja punto com* [B.11], y será a mediados del año 2001 cuando finalmente estalle, generando una crisis de confianza y marcando un momento crucial para la Web. Este estallido provocó que el siguiente ciclo de evolución tuviera un ritmo de crecimiento menos arriesgado, favoreciendo así su consolidación.

La madurez de la sociedad, el contacto con las nuevas tecnologías y la digitalización e intercambio masivo de contenido multimedia (imágenes, música y vídeo) generaron una nueva forma de interactuar donde no solo se busca contenidos, sino que también se aportan. La evolución de los formatos en que se desarrollaban las páginas Web permitió la separación entre los contenidos y la manera en que se visualizaban por pantalla, facilitando así enormemente la exportación de contenidos de diferentes fuentes y la integración de los mismos en nuevos contenedores.

Esta realidad generalizada y en respuesta al contexto socio-económico en crisis que se vivía, cerró un capítulo de la historia de Internet, la Web 1.0, dando a luz al paradigma actual, la Web 2.0[1].

De este modo, el término Web 2.0 se puede definir como el grupo de aplicaciones Web que fomentan la comunicación participativa por todos aquellos usuarios que la utilizan o visitan, agregando nuevos contenidos de distintas fuentes o formatos por parte de estos. Está forma de interacción cambia radicalmente el rol del usuario, cambiando de un perfil limitado a la observación pasiva de los contenidos, a uno cuya finalidad es ser productores y consumidores de los mismos. Desde este momento, la Web ya no es lo que era antes.

	Web 1.0	Web 1.5	Web 2.0
Tipo	Estática	Dinámica	Colaborativa
Fechas	1994 a 1997	1997 a 2003	2003 a actualmente
Tecnología	HTML	SSL, DHTML, CSS, y las anteriores	AJAX, XML, y las anteriores
Descripción	Contenidos poco actualizados, y generados por el propietario de la página Web	Contenidos actualizados por los usuarios a través de editores de contenidos	Contenidos completamente colaborativos

Figura 2. Tabla comparativa de las distintas versiones Web

Se puede observar que más que presentar novedosas tecnologías, el término Web 2.0 revela una nueva forma de uso de la Web, en las que se definen por estar en continua evolución y actualización, llegando a decirse que se encuentran en un estado denominado beta perpetuo (en inglés, *permanent beta*). Estas características son presentadas por Tim O'Reilly en su artículo [B.12], donde expone siete cualidades de lo que se entiende como Web 2.0.

- La Web a modo de plataforma.
- La inteligencia colectiva.
- La gestión de las bases de datos.
- Los modelos de programación.
- El final del período de las actualizaciones del software.
- El software multiplataforma.
- La práctica enriquecedora de los usuarios.

Estos atributos difieren bastante de lo que se considera como Web 1.0. La actual Web para a ser un medio de trabajo que no necesita de un software instalado en la maquina local.

1. Término definido por Dale Dougherty de O'Reilly Media en una tormenta de ideas con Craig Cline de MediaLive para la conferencia organizada en Octubre del 2004, bajo el lema *La Web como plataforma*.

Actualmente, todo está en la Web, más que nunca. La Web se convierte en el sistema donde se ejecuta el software, teniendo la oportunidad de usarlo sin tener que instalar ningún software en el equipo. De esta forma, se superan algunas de las barreras de tipo tecnológico, y el concepto de Escritorio (en inglés, *Desktop*) evoluciona a Escritorio Web (en inglés, *Webtop*).

La inteligencia colectiva es considerada como la energía de la Web 2.0. La aparición de las nuevas herramientas en las que todos los usuarios son capaces de publicar y comentar, está generando información de interés que puede ser usada por otros. Como ejemplo paradigmático de esto, encontramos la Wikipedia. Sin embargo, bajo este paradigma de participación existen problemas de verificación de contenidos, pues no todo lo que se ha publicado es por especialistas del tema, sino más bien por la cooperación, la creatividad y las críticas de los usuarios.

Por otro lado, esta inteligencia colectiva que conforma a la Web participativa, es aprovechada para los negocios de muchas empresas que quieren conocer las características y prioridades de los usuarios mediante instrumentos que consumen información de Internet.

Toda esta inteligencia colectiva y creación de novedosos contenidos, está creando grandiosos volúmenes de información dispersos por Internet. Es por ello, que uno de los puntos clave de competencia esencial de las compañías actuales es la gestión de la base de datos. El control sobre estas supone añadir un valor agregado a los datos que la contiene, y por ende, un valor comercial a los productos de las empresas. Se debe pensar en herramientas 2.0 como una arquitectura de tres pisos: la comunidad, el negocio y la tecnología.

El desarrollo de estas herramientas 2.0 apuesta por la programación ligera donde se admite la integración de asociaciones no existentes en los paquetes informáticos, proporcionando así simplicidad a que el usuario logre acceder y participar a los contenidos cuando él lo desee de una forma más fácil y sencilla.

Una de las características importantes de estas herramientas es su dinamismo. La evolución y generación de nuevos contenidos cambian casi a diario, debido principalmente al papel que juega el usuario en estos sistemas. No es solo un colaborador que aporta información al sistema, sino que también se encargará de decidir cuales son las funcionalidades que deben continuar y cuales no. Es por tanto, que las herramientas 2.0 deben de desarrollarse diariamente siguiendo las reglas que va marcando el usuario.

Estas herramientas no son solo propietarias de un único dispositivo (ordenador). La aparición de las nuevas tecnologías, los nuevos dispositivos digitales y la sociedad de consumo en la que vivimos, esta favoreciendo que cada vez existan más formas en las que interactuar con Internet. Y es por ello, que la Web 2.0 se amplía a más dispositivos, y con esto, a una mayor cobertura para las herramientas 2.0, ofreciendo al usuario un acceso amigable desde cualquier lugar y momento.

Lejos han quedado las tradicionales apariencias visuales en las que se brindaba la sobrecarga de objetos animados. La Web 2.0 apuesta por la apariencia interactiva con el usuario, de tal forma que pueda moverse y operar de la misma manera como en las aplicaciones locales, garantizando un entorno amigable como condición de usabilidad. Para que una aplicación 2.0 tenga éxito, debe de aprender de las exigencias de interacción del usuario sin dejar de proporcionar servicios de colaboración y creación de datos compartidos.

2.2 Evolución hacia la nueva sociedad

Como bien plasmó Charles Darwin [B.13] en su teoría de la evolución, para que el ser humano pueda sobrevivir como especie se ha de acoplar a los cambios que van surgiendo, buscando la mejor solución para abordarlos. Con esta teoría, el ser humano creó el lenguaje como sistema de comunicación, cambiando así la forma de comunicarse con el resto de personas. La evolución de la ciencia y la tecnología generaron nuevos sistemas de comunicación, nuevas formas de relacionarse y comunicarse, el telégrafo, el teléfono, la televisión...hasta llegar a nuestros días, donde gran parte de esta comunicación se realiza a través de los

medios digitales conectados entre sí gracias a la red de Internet.

Inicialmente, en la llamada sociedad 1.0 o sociedad real, los usuarios exclusivamente podían consultar la información existente sin poder aportar alguna colaboración. En este momento, el usuario se encontraba oculto en un segundo plano y los métodos de comunicación únicamente se basan en la captación de información expuesta por el creador del sitio Web.

La primera evolución fue ocasionada por la transformación de la Web a una plataforma de lectura y escritura, Internet sufre un cambio paradigmático en su interior, donde la información ya no se está generando por unos usuarios o servicios en concreto, sino que dichos contenidos están siendo creados por usuarios corrientes de la Web. Esta revolución introdujo nuevos conceptos como los W*eblogs*, las primeras *wikis* [B.14] y los diarios personales. En ese momento, un usuario no solo podía consultar información, sino que también podía editarla de forma rápida y fácil.

La tradicional Web 1.0, donde la información es cerrada y el usuario queda al margen de la participación sobre esta, pasa a ser una nueva Web elaborada por los usuarios implicados en la colaboración e interacción, donde se rompe las barreras de uso y creación exclusivos para unos determinados usuarios o servicios.

Afortunadamente esta transición generó una evolución hacía una realidad completamente distinta a la conocida hasta el momento, donde Internet comenzó a estar repleto de medios en los que los usuarios se podían dar a conocer a la sociedad.

De este cambio surge la llamada Web 2.0, la cual se considera como una manera de modificar los contenidos de Internet de tal forma que cualquier usuario, consiga tener una experiencia completa de la red. Este proceso aparece por una serie de fenómenos polifacéticos, como pueden ser los *blogs*, los servicios *online*, las redes sociales o el amplio mundo de servicios y nuevos usos sociales que se generan a su alrededor.

La Web 2.0 no solo aporta nuevas herramientas tecnológicas con la que los usuarios pueden ser más colaborativos y participativos, sino que enmarca una serie de fenómenos socio-culturales producidos por la necesidad de expresión por parte del usuario y la crisis socio-económica vivida. A esta nueva sociedad se la conoce como Sociedad 2.0 y está compuesta por todos aquellos usuarios que colaboran de forma activa en la red.

La Web 2.0 se constituye por el dinamismo y la participación, así pues, se pueden encontrar aplicaciones que facilitan a los usuarios la publicación de información en la red, convirtiéndose estos en consumidores y productores de la información. La adquisición de las características propias de un *software social* [B.15]*:* la comunicación, que permite poner conocimiento en común, la comunidad, que ayuda al ingreso de los usuarios, y la cooperación, que ayuda a que los usuarios logren los objetivos, esta provocando que la Web 2.0 actúe como lugares sociales donde se reúnen los usuarios.

Así pues, gracias a la integración del *software social* en el modelo de Web 2.0, a su continua expansión (ya que cada vez más aparecen nuevas experiencias y su demanda se ve aumentada por los usuarios), y al crecimiento exponencial de usuarios que acceden y utilizan estos servicios, ha provocado una nueva revolución social, en la que miles de millones de personas se relacionan compartiendo información a través de este medio, llegando a afectar en muchos de los aspectos sociales de estos. Este cambio o revolución es conocido como la democratización de Internet.

Los usuarios son generadores y productores de sus intereses. Mediante la utilización de buscadores, estos pueden localizar fácilmente sitios especializados en su tema de interés, permitiéndoles captar información, visualizar contenidos y compartir sus inquietudes. Así pues, esta libertad está provocando que los usuarios fuercen a que se hagan productos de mejor calidad y que se hagan más competitivos. Y es por ello que las empresas están cambiando su visión organizativa para adaptarse a esta revolución. La interacción con

el cliente mediante las nuevas vías de comunicación, es actualmente un factor clave de evolución para la empresa.

El paso de una Web oculta a una participativa (Web 2.0) se aproxima al ideal de Internet desde el punto de vista de la interacción social. Esto se hace evidente cuando son los usuarios los que aportan contenidos a la Web, favoreciendo así la inteligencia colectiva.

A pesar de todo, aún no son muchos los que no son conscientes del verdadero poder de este tipo de Webs. Esto queda demostrado por la regla del 1% [B.16], donde se expone que cada cien usuarios que entran a un portal, el 90% únicamente busca información, el 9% participan y únicamente el 1% aportará nuevos contenidos.

Desde un punto de vista psicológico y social, a Internet se le atribuye innumerables beneficios para la educación, el mundo empresarial, el social o el personal. El uso de las Web 2.0 mejora las funciones psicológicas y proporciona una mayor alfabetización visual. Sin embargo, el exceso de este puede influenciar negativamente a los usuarios [B.17]:

- Menor esfuerzo mental.
- Poca atención a la información verbal.
- Disminución de la imaginación.
- Reducción del círculo social.
- Remplazo de las uniones cercanas.
- …

Con la llegada de Internet y su evolución imparable, algunas corrientes psicológicas piensan que han surgido nuevas afecciones psicopatológicas relacionadas con este medio. Es evidente que Internet está cambiando la cultura y los hábitos de las personas, el uso excesivo de este, está provocando un descenso radical de la comunicación entre los individuos, y en consecuencia, está disminuyendo el círculo social y aumentado la sensación de soledad e individualismo. El exceso de uso de Internet está provocando un aumento en los niveles de depresión, una depresión cíclica que alimenta el uso del ordenar e Internet, y por tanto, el aislamiento y la independencia.

En la educación, según algunos psicólogos, Internet ha favorecido a la disminución de las capacidades mentales. Gracias a este y a la cantidad de información que contiene, ha provocado el deterioro de la capacidad de memorizar y acumular conocimiento. Y esto se ve agravado aún más en los niños y adolescentes, donde en su etapa de crecimiento deberían de explotar su capacidad de memorizar y almacenar todo lo que aprenden, y utilizar su cerebro para generar nuevos conocimientos. Todo esto, está favoreciendo a que la actual sociedad tenga problemas de comprensión y grandes deficiencias a la hora de razonar.

No obstante, Internet puede favorecer la enseñanza en muchos aspectos. La capacidad de obtener información, las posibilidades de comunicación entre las personas, las actividades cooperativas, el conocimiento de las nuevas tecnologías, la motivación e interés por conocer, las habilidades de selección y búsqueda de información…son algunos de los muchos factores por los que Internet es uno de los medios más utilizados.

Debido a la era en la que la sociedad vive actualmente es difícil no utilizar las nuevas tecnologías, y más bien, sería algo contradictorio debido a las ventajas que ofrece, pero siempre y como todo, con un uso responsable y adecuado.

2.3 Contexto

Como en toda evolución, la Web ha seguido una ruta lógica de progreso en la que ha sufrido constantes cambios hasta llegar a ser lo que conocemos actualmente.

La Web original o 1.0, es la forma más elemental que se conoce. Por aquel entonces, todo estaba formado por texto plano, y existían unos pocos navegadores que eran capaces de interpretarlo. Con la aparición del HTML como lenguaje de hipertexto, se posibilitó la integración de componentes multimedia dándole un aspecto visual más llamativo. Con ello, surgieron las primeras páginas Web estáticas, en las que su elevado contenido se centraba en la reunión de elementos multimedia. La poca constancia de actualización y de introducción de nuevos contenidos, provocó una pequeña evolución de la Web, donde aparecieron las primeras páginas dinámicas.

Esta aproximación a lo que se conoce actualmente como Web originó la aparición de unos sistemas de interacción con el usuario que servían como complemento de un sitio Web. A estos nuevos servicios se les conoce como foros.

2.3.1 Foro

Los foros surgieron por la necesidad de una aproximación e interacción con los usuarios. Estos invitan al usuario a introducirse en una discusión libre e informal sobre un tema y así generar nuevos contenidos, y por tanto, nueva información. Con la aparición de diversos usuarios discutiendo y opinando sobre un mismo tema, surge un nuevo concepto en Internet, la comunidad virtual. Esto puede ser visto como un tablón de anuncios en el que cada usuario puede dejar sus impresiones sobre un determinado tema.

Estas plataformas continúan siendo una de las herramientas principales de intercambio de información y opiniones, aunque en muchas ocasiones han sido tachados de provocadores del alboroto y el ruido en Internet. Esto es debido principalmente al mal uso que se les ha dado y al poco respeto por parte de los usuarios participativos.

Los foros por tanto, son entendidos como las redes sociales del pasado, donde el usuario comenzó a exponer sus inquietudes y a ser más interactivo con la Web, pero siempre limitado a las aportaciones que presentaba el *webmaster* como inicio del debate y supervisado en todo momento por este.

2.3.2 Anotación social

A consecuencia de estas limitaciones, a la necesidad de comunicación del usuario, a la búsqueda de nuevas experiencias y a la evolución de la Web, surgen nuevas aplicaciones asociadas al fenómeno social de la participación y la interacción, concebidas para un uso participativo, donde el usuario puede clasificar, valorar, crear, actualizar, modificar y eliminar nuevos contenidos.

De este surgimiento, nacen las primeras aplicaciones colaborativas orientadas a la creación, las *wikis*, sitios Web donde los contenidos de la página pueden ser generados, modificados o eliminados por múltiples usuarios a través del navegador Web. Uno de los principales usos que se les ha dado a este tipo de páginas Web ha sido la creación de enciclopedias colaborativas, siendo una de las más potentes en esta rama la Wikipedia [E.1].

De la falta de clasificación jerarquizada de la anterior Web, surge la idea de utilizar etiquetas (del inglés, *tags*) para clasificar los contenidos de una manera informal. Los usuarios son participes de este fenómeno conocido como folksonomía [B.18] cuya finalidad es adjuntar etiquetas en un espacio de nombres, sin categorías ni relaciones de parentesco establecidos, a los contenidos publicados.

Como resultado de este proceso, se hace posible la ordenación y clasificación de los contenidos. Usualmente estas etiquetas son mostradas en forma de nube de etiquetas (del inglés, *tag clouds*) en las que se agrupan visualmente las etiquetas en diferentes tamaños, teniendo los más populares un tamaño mayor a los demás. Estas etiquetas son utilizadas por aplicaciones compartidas como del.icio.us [E.2] o buscadores específicos como *Technorati* [E.3].

Además, en ayuda a esta jerarquización y clasificación, surgen las valoraciones o priorizaciones de contenidos Web, donde el usuario no sólo puede aportar sino que también valorar, en base a un criterio que normalmente suele aparecer en la página principal. Algunos ejemplos de este tipo de aplicación son *Tripadvisor* [E.4], *Digg* [E.5] o *Aupatu* [E.6].

Figura 3. Anotación social: Wiki y valoración

2.3.3 Blog

Con este progreso enmarcada por la capacidad de aportar comentarios y valorar los contenidos por parte de los visitantes del sitio Web, enriqueciendo la página y conformando una comunidad en torno al mismo, nacen los *blogs*. Este término proviene de contraer W*eb-log*, también conocido como bitácora, ciberdiario, cuaderno de bitácora o ciberbitácora [B.19]. Básicamente es un sitio Web con aportaciones periódicas de uno o varios usuarios, y en que las entradas están datadas y ordenadas cronológicamente, mostrando al inicio el contenido más reciente.

La popularidad de este tipo de páginas Web esta ligado fuertemente a la creación de herramientas que facilitan considerablemente la generación y mantenimiento de un *blog* a todas las personas incluso aquellas sin experiencia en informática. Actualmente disponemos de diversas aplicaciones para hacer uso de este tipo de páginas, de las cuales hay que destacar *Wordpress* [E.7] y *Blogger* [E.8] por ser las más conocidas.

Gracias a las nuevas tecnologías, actualmente, en los blogs no solo se comparte experiencia plasmada en texto plano, sino que son capaces de compartir ficheros multimedia, tanto imágenes estáticas como vídeos. A estos últimos *blogs* se los conoce como *videoblogs*, donde los usuarios pueden aportar contenido audiovisual. Para este tipo de uso, encontramos herramientas tales como: *Blipback* [E.9].

La utilización de estos *blogs* como seguimiento de la actividad diaria realizada por una persona ha generado la aparición del fenómeno conocido como *microblogging*. A través de aplicaciones Web, y textos cortos (aproximadamente el tamaño de un mensaje de texto telefónico), los usuarios mantienen un registro actualizado de las acciones u opiniones en las que se encuentra inmerso. Uno de los más conocidos es: *Twitter* [E.10].

Figura 4. Blogs

2.3.4 RSS

Debido al gran éxito de estas plataformas y a su continua evolución de contenidos, surgió un mecanismo por el cual un usuario puede estar suscrito al canal de la página Web y así recibir las últimas actualizaciones de los contenidos, ofreciendo al usuario la posibilidad de estar siempre enterado de las últimas novedades que surjan en esta.

Los RSS (del inglés, *Really Simple Syndication*) son uno de las columnas principales de la Web 2.0. Usualmente conocido como sindicación, este sistema sencillo consiste en la distribución de los contenidos de una página Web mediante un formato estandarizado (XML, del inglés *Extensible Markup Language*). El sistema permite distribuir los contenidos sin necesidad de un navegador, de manera que cualquier usuario pueda suscribirse a este, agregando el canal (*feed*) a su programa de lectura de contenidos RSS (agregador). En las más actuales versiones de los navegadores más conocidos, es posible la suscripción a estos canales sin la necesidad de utilizar un programa adicional para su lectura.

Algunos de las aplicaciones para RSS más utilizadas son: *Google Reader* [E.11] o *Blogbridge* [E.12].

Figura 5. RSS

2.3.5 Geolocalización

Siguiendo con la evolución y la necesidad de conocer *el todo* de la información, surgió la georreferenciación, como respuesta a donde se originó el contenido. Este tipo de aplicación es uno de los sistemas con más auge dentro de las herramientas 2.0. Gracias a su aparición en los últimos años, y principalmente a su uso extremadamente sencillo, el usuario es capaz de georreferenciar cualquier contenido o elemento presente en el mundo e integrar sus coordenadas (posición) en un entorno Web.

El uso de herramientas como *GooleMaps* [E.13] o *Yahoo! Maps* [E.14] ha generado un impacto sociológico, apresurando la creación de las Web geosemánticas, cuya arquitectura de servicios Web esta basada en ontologías, y esta diseñada para integrar y compartir información variada, permitiendo la transferencia y unificación de datos geográficos entre bases de datos.

Figura 6. Geolocalización

2.3.6 Redes sociales

De la constante evolución de estos canales de comunicación, de los cambios socio-culturales sufridos, de la aún más necesidad de compartir y generar información entre los usuarios, de las comunidades virtuales y las relaciones entre los usuario, surgen las redes sociales online.

Este término no es para nada nuevo, ni su origen proviene del nacimiento de la Web 2.0. Sin embargo, su llegada ha provocado el despegue a una velocidad vertiginosa, debido a la necesidad de socialización por parte de los individuos.

Si evaluamos estos sistemas desde una visión antropológica, estos nacen tras la Segunda Guerra Mundial, al surgir la necesidad de entender las conductas de los individuos al adaptarse a un nuevo entorno y apartarse de los modelos culturales estáticos. Estos estudios son utilizados para identificar las estructuras sociales salientes de las relaciones, así como un conjunto de técnicas y métodos relacionados con estas.

Desde un punto de vista sociológico estas redes se utilizan como herramienta de consolidación de la sociedad, potenciando sus características integradoras como la horizontalidad, la creatividad y la solidaridad. Para esta corriente, las redes sociales son técnicas representativas donde se percibe a los grupos de individuos como complejos sistemas de comunicación e intercambio a lo largo de enlaces interconectados. Las relaciones entre la generación y el consumo de información están cambiando rápidamente, la gestión de la información individualizada y la aparición de una nueva argumentación de democratización, provoca que los usuarios demanden de Internet su propio espacio donde controlar sus contenidos.

Estos modelos colaborativos giran en torno a una nueva cultura donde cualquier usuario puede participar, y todos tienen la posibilidad de ser vistos y escuchados. La construcción de conocimiento propio, de los demás o del mundo entero, va generando grafos de información de cómo las personas se observan, se comunican, se relacionan, comparten... Esto está consolidando el término de Web 2.0, donde los usuarios se van involucrando en la creación de contenidos de la Web al mismo tiempo que consumen estos.

Con esto en mente, las redes sociales son estructuras sociales alimentadas por personas, facilitando la conexión entre ellas a través del registro de las relaciones que cada miembro mantiene en diferentes entornos, tanto para fines empresariales como personales.

Actualmente existen infinidad de redes sociales, algunas de las más comunes son: Twitter, Facebook [E.15], Linkedin [E.16], Orkut [E.17], o Xing [E.18].

Figura 7. Redes sociales

2.3.7 Aplicaciones sociales

Gracias a los avances tecnológicos y a las nuevas plataformas, el usuario posee la capacidad de acceder a la información en cualquier lugar y en cualquier momento. Pero la Web 2.0 pretende llegar aún más lejos, a un lugar donde la información no tenga que ser almacenada por los usuarios, sino que se encuentre alojada en servidores externos, un sitio donde los usuarios puedan hacer uso de herramientas sin tener la necesidad de ser instaladas en las máquinas locales, un área donde los usuarios colaboren en conjunto con las empresas…

Dentro de los usos de la Web 2.0 y más concreto en la capacidad de crear e intercambiar contenidos a través de la Web, han surgido aplicaciones cuya finalidad es la de almacenar recursos en Internet para poderlos visualizar y compartir con el resto de usuarios. La mayor parte de estos servicios son conocidos por los usuarios, como es el caso de Flickr [E.19], YouTube [E.20], Slideshare [E.21], GoogleDocs [E.22].

Figura 8. Intercambio de contenidos

2.4 La información en el caos

Con la aparición de Internet los hábitos de las personas están cambiando de una forma sorprendente. El uso de esta tecnología junto con los medios digitales está transformando la cultura en la que viven, las costumbres habituales y muchas de las disciplinas a las que estaban acostumbrados. A través de estas tecnologías, las personas reciben grandiosas cantidades de información que en muchos de los casos no se dominan, por lo que pueden provocar inquietudes e incertidumbres en los usuarios.

Si analizamos el significado del término información en la teoría de la información, esta se define como la agrupación ordenada de datos analizados que cambian el estado de conocimiento de una persona. Como se puede apreciar, este término expresa claramente una organización y ordenación. A pesar de los elementos de la teoría de la información: fuente, mensaje, código e información, existe otro como la entropía.

Según la teoría de la información [B.20], se habla de entropía a la determinación de la parte energética que no puede ser usada para producir trabajo. La información es caracterizada como una ordenación de símbolos físicos mediante este término. Considerando que los canales de información no son los más adecuados, la entropía estima en mayor medida la probabilidad de la cantidad de información útil que puede ser transferida a través del medio. Así pues, la entropía es considerada como una medida de la incertidumbre, que a partir de un desorden puede generar un orden.

En ocasiones el uso de energía puede generar órdenes, pero en otras, el destino origina incertidumbre en las interpretaciones, llegando a un estado de caos, o desde otro punto de vista, a la generación de nuevos órdenes dentro del caos.

Internet, las páginas Web y sus contenidos progresan y se modifican diariamente, cambiando su forma y en algunos casos su existencia. A su vez, en las páginas Web los enlaces se reproducen, los usuarios aumentan exponencialmente...provocando nuevas formas de comunicación, mayores posibilidades y más cantidades de información para ser consultadas. Es un medio en que la información y los datos existen en abundancia y se va de la incertidumbre a lo previsible, por lo que, existe una relación entre el orden y el caos.

Internet se comporta como un amasijo de autopistas cibernéticas, que forman un caos donde el usuario a de conducir bajo estos efectos. La circulación por estas autopistas se realiza mediante enlaces y tienen como objetivo guiar al usuario en las posibles soluciones a las necesidades de información de un usuario. Estos son las conexiones entre las autopistas, son formas en que el usuario puede desplazarse a través de la red. Las experiencias de los usuarios en los recorridos por los enlaces permiten elaborar nuevos órdenes.

Los recuperadores de información se comportan como generadores del orden de estas autopistas, donde los usuarios realizan consultas de información en este espacio para obtener unos resultados ordenados.

Estos recuperadores se encargan de generar nuevos órdenes a partir del caos de miles de millones de documentos desperdigados en diferentes puntos de esta red de autopistas. Algunos de ellos, ofrecen la posibilidad al usuario de configurar estos órdenes, facilitándole la interpretación y análisis de la información obtenida a partir del caos.

Los usuarios que circulan por estas autopistas, constantemente están expuestos a prácticas entrópicas, debido principalmente a la enorme cantidad de información existente. Y es que, Internet ofrece infinidad de posibilidades para los usuarios, donde los caminos proporcionan nuevas experiencias para estos, y que en algunos casos producen incertidumbres. No obstante, la utilización de conceptos cotidianos se ven reflejadas en los sistemas digitales. El usuario navega por estas autopistas mediante el uso de tecnología, pero desde una visión de lo usual y de lo que está acostumbrado. Los términos como ventanas, navegación, portales...son algunos de estas expresiones del lenguaje común, y que se utilizan en el lenguaje virtual.

Por tanto, el caos de Internet no hay que concebirlo como un amasijo de autopistas o canales de información que inducen a la incertidumbre, sino como un medio que da paso a un nuevo orden, un orden lleno de posibilidades para los usuarios y las empresas, un orden donde la información está esperando a ser recogida.

Y aquí es donde la escucha activa tiene un papel crucial.

3. ESTADO DEL ARTE

3.1 Introducción

Desde sus orígenes, el ser humano ha sentido, siente y sentirá una necesidad de representar y dejar constancia de todo lo que le rodea. Inicialmente se utilizaba la escritura, la cual ha sido el sistema tradicional para reflejar el conocimiento, pero gracias a los avances tecnológicos de los últimos sesenta años, han ido surgiendo diferentes medios de representación de información donde esta es mostrada y almacenada digitalmente, facilitando la distribución de forma sencilla y masiva por medio de redes de computadores.

La recuperación de información no se centra solo en el paradigma de la tecnología, el ser humano está captando información constantemente. La resolución a un problema, la documentación de una tesis… son actividades que conllevan una recuperación de información. Con la evolución de las nuevas tecnologías y las nuevas tendencias sociales, las técnicas de recuperación de información se encuentran en cualquier actividad en entornos digitales.

La aparición de esta nueva generación de tecnologías de la información y comunicaciones [B.21] (TIC, en inglés *Information and Communications Technology*) han generado una nueva era donde el conocimiento y la información es uno de los activos intangibles más apreciados para las organizaciones, por lo que se precisan de herramientas que consigan captar el mayor número de información de Internet. Estos sistemas deben de realizar la tarea de recopilar, analizar, organizar y diseminar toda la información posible para que en un momento dado se pueda brindar la información adecuada a la organización aportándole un valor añadido.

Como hemos visto en la introducción, uno de los medios más importantes en esta evolución es Internet, el cual es el sistema de generación y transmisión de información más grande conocido. Las nuevas vías de comunicación dentro de este ámbito, entre ellas las herramientas colaborativas que permiten a cualquier persona compartir información: *blogs*, foros, redes sociales... están generando un volumen de información extremadamente grande y desestructurada, cuyo crecimiento se ve aumentado exponencialmente cada año. Esta información no solo se encuentra en forma textual, cuya representación es el más simple, sino que esta puede adquirir diferentes formas como son los vídeos, la música, las librerías, los espacios Web... No obstante, debido a este crecimiento, su problema principal es el descontrol de la información a la que los usuarios pueden acceder, provocando dificultades para encontrar de manera simple y eficiente información de calidad, y afectando a la usabilidad de los sistemas de recuperación.

En consecuencia a este crecimiento, ha surgido un nuevo fenómeno, el cual es denominado por algunos investigadores [B.22][B.23] como la sobrecarga informativa (IO en inglés, I*nformation Overload*). Este concepto es utilizado cuando un usuario se encuentra en una situación en la que tiene más información de la que es capaz de procesar, y por consecuencia, puede llevar a realizar acciones equivocas o dejar de tomar alguna de ellas. Por ello, uno de los puntos clave de todo sistema de recuperación de información en la Web, es conocer la formula idónea para localizar la información disponible en Internet, y mostrársela al usuario de forma que pueda ser procesada por este.

Actualmente, los sistemas de recuperación de información en la Web como los buscadores, se han convertido en los principales mecanismos de interacción con los usuarios que necesitan obtener información, llegando a ser una actividad frecuente realizada en el día a día. No obstante, la aparición de herramientas como las redes sociales y los medios de intercambio de contenidos, se han convertido en otros nichos de información, que a través de sus sistemas de interrogación proporcionan datos de relevancia a los sistemas de consultas.

Desde el punto de vista de un sistema de recuperación de información en la Web, estos escenarios consisten en servicios que atendiendo a los criterios de la consulta y mediante el acceso a sus bases de datos, ofrecen al sistema, información de interés para resolver las necesidades del usuario. Pero además de entregar dicha información, se facilita al usuario la navegación entre dicha información, evitando que este tenga que verse inmerso en la red de autopistas de Internet. Desde un punto de vista de abstracción, estos sistemas recuperan toda la información, pero a su vez, restringen aquellos recursos que no cumplen con los requisitos deseados establecidos por el usuario.

Llegados a este punto, se hace preciso realizar las siguientes preguntas.... ¿Y para que tanta información?, ¿Que necesidades hay para querer obtenerla?

Como se ha visto anteriormente, actualmente Internet es el medio más grande de información donde muchas personas expresan sus inquietudes y opiniones sobre temas concretos. Debido a esto, y desde un punto de vista de las organizaciones, captar toda esta información y analizarla se convierte en un proceso fundamental para obtener resultados óptimos en sus campañas. El estudio de su ecosistema social en la red a través de los resultados obtenidos por la herramienta de recuperación de información, resulta interesante para fidelizar a los usuarios, atendiendo sus preocupaciones o intereses, escuchando sus propuestas o solicitudes, e involucrándolos con los objetivos de la organización. Por todo ello, y debido a la era de información en la que vivimos, se hace indispensable la utilización de este tipo de herramientas para conocer el impacto social de la estrategia seguida por una organización y sus consecuencias que suscita en los usuarios.

Bajo estos pilares, se ha desarrollado el sistema de recuperación de información como módulo indispensable de la escucha activa en Internet. No obstante, se hace recomendable hacer un recorrido de donde proviene la recuperación de información, cuales son sus modelos, que evolución a seguido en concordancia con los actuales medios de información, cuales son los principales sistemas de recuperación de información en Internet, donde se encuentra la información y en que medida, cual es la importancia de esta información y que puede aportar a las empresas actuales.

3.2 Información

Durante todo el documento hemos visto y veremos mencionado el término información, cual es su importancia y la gran cantidad a la que nos enfrentamos actualmente. Pero, ¿Qué es la información?.

Hoy en día resulta paradójico la dificultad que presenta encontrar una única definición para esta palabra, viéndose arraigada por la era en la que vivimos y que día a día van surgiendo herramientas en las que el usuario puede manipularla.

A grandes rasgos, la información [B.24] se puede definir como el conjunto organizado de conocimientos que se extrae a partir de un dato procesado, cuyo objetivo es reducir la incertidumbre a la persona o sistema que recibe el mensaje.

Según la pirámide de la información, un dato ha de sobrepasar dos niveles hasta convertirse en conocimiento para la persona. Los datos son la unidad principal de la evolución piramidal, a partir de ellos se construye el resto de atributos, y mediante su análisis y procesamiento es cuando se convierten en información. Si además, el usuario asimila dicha información, adquiere un saber o conocimiento que le será útil en una situación concreta.

Figura 9. Pirámide de la información

3.3 Recuperación de información

Con la premisa en mente de lo que es la información y como puede ser convertida en conocimiento, veamos pues en que consiste la recuperación de información y cuales fueron sus orígenes.

La recuperación de información (RI, del inglés *Information Retriveal*) [B.25] no es un campo nuevo de investigación. En el periodo de 1950 a 1960 surgieron los primeros sistemas automáticos de recuperación de información, sin embargo, debido a la importancia que tiene la información actualmente, estos sistemas son una herramienta de gran utilidad.

El refuerzo de la plataforma Web como sistema de información, ha influido en el desarrollo de sistemas de recuperación de información (SRI, del inglés *System Information Retriveal*) más avanzados, donde el objetivo de estos es la identificación de una o más reseñas Web que resulten importantes para satisfacer una necesidad de información por parte del usuario. Se puede plantear que disponer de la información más adecuada en un momento determinado puede resultar el éxito o fracaso de una acción. De esto, se deduce la importancia de los SRI a la hora de manejar estas situaciones de manera eficaz y eficiente.

Actualmente se pueden encontrar varias definiciones para la recuperación de información, resultando un poco difícil establecerlas dentro del campo de las ciencias de información. Cabe destacar algunas propuestas como la de Salton [B.26] : "*es un campo relacionado con la representación, alojamiento, ordenamiento y acceso a los elementos de información*", o la de Croft [B.27], el cual define que la RI es "*son las tareas por las cuales el usuario identifica y accede a los elementos de información que son adecuados para la resolución del problema dado. Dentro de estas tareas los lenguajes documentales, las técnicas de resumen, la descripción del objeto documental... son de gran importancia*", o Meadow [B.28], el cual detalla que la RI es "*un conjunto de reglas mediante las cuales se localiza una determinada información dentro de una base de datos o almacén*".

Desde los últimos años [B.29] se están realizando estudios sobre la interacción de la Web y los sistemas de la Inteligencia Artificial para mejorar el acceso a la información, obteniéndose unos resultados atractivos y de notable éxito. De esta interacción surge lo que se denomina la Web Inteligente o Web 3.0 [B.30] (WI, del inglés *Intelligence Web*) cuyo objetivo es la de ofrecer servicios más inteligentes a los

usuarios. Dentro de este término nos encontramos con conceptos como la Web Semántica (del inglés *Semantic Web*), la Web Geoespacial (del inglés *Geospatial Web*) o la Web 3D.

Cuando un usuario siente la necesidad de realizar una consulta para obtener información, comienza a suscitar una serie de acciones que conllevan el devolver el equilibrio a su estado inicial. Estas acciones suelen estar medidas por un proceso de comunicación, que en nuestro caso es mediante la interacción del usuario con una máquina. No obstante, en ocasiones la resolución a esta necesidad de información, puede ocasionar resultados que provoquen nuevos problemas.

Por todo lo comentado anteriormente, se puede concluir que un sistema de recuperación de información es un tipo de sistema de información que trabaja con bases de datos formadas por documentos (local u *online*), mediante las cuales se localiza, obtiene, procesa y almacena la información relevante a partir de una petición establecida por el usuario en una sentencia basada en lenguajes de consultas. Así pues, todo sistema de recuperación de información esta constituido por las siguientes premisas:

- Se tiene una colección de documentos que contienen información de interés para un usuario.
- Hay al menos un usuario con la necesidad de información sobre un tema en concreto.
- Este usuario realiza una petición al sistema de recuperación de información.
- En respuesta a la consulta realizada por el usuario, el sistema retorna una colección de documentos con información relevante para este.
- Por último se presentan los resultados mediante una interfaz gráfica.

Las nuevas tecnologías han favorecido que la mayor parte de la información se encuentre digitalizada y distribuida por la red, además muchas de las actividades de búsqueda y localización se realicen sobre este medio. Es por ello, que la recuperación de información en Internet es una de las tareas más importantes actualmente y de mayor auge. No obstante, se hace necesario ver los modelos que tienen este sistema para trabajar con documentos locales.

3.3.1 Modelos tradicionales

Desde los inicios de la recuperación de información se han ido desarrollando métodos que faciliten la comparación de una consulta con la información disponible en la base de conocimiento. A continuación, se describirán de forma breve cada uno de ellos:

- El modelo Booleano, se considera una recuperación sencilla basada en el álgebra booleana y en la teoría de los conjuntos. Debido a su peculiar simplicidad, ha sido adoptado por muchos de los sistemas de recuperación de información. Su sistema se basa en determinar si un documento es importante o no mediante un criterio de decisión binaria. Sin embargo, por su sencillez acarrea muchos inconvenientes de los cuales se puede destacar el no tener en cuenta el número de apariciones de una palabra en el documento o el no calcular el grado de relevancia para los documentos que componen la respuesta.

- El modelo probabilístico, se centra en la distribución estadística de los términos para mejorar el rendimiento del sistema, de manera que la aparición de uno de los términos en un documento o varios es considerado como un parámetro relevante de similitud o no. Así pues el sistema, puede calcular el grado de similitud entre la consulta y los documentos, y mostrar los resultados según un orden de relevancia.

- El modelo vectorial, propone un marco en el cual sea posible la coincidencia parcial en el valor de relevancia de los documentos. Se crean grupos de documentos similares, partiendo de una representación vectorial de estos, quedando entonces, en un lugar determinado por su posición. Actualmente, es uno de los métodos más utilizados, debido principalmente por el esquema que utiliza, a su equiparación parcial y a la ponderación de los términos tanto en la consulta como en el documento.

- El modelo estructurado, mediante la estructura de los documentos el sistema evalúa la relevancia de cada uno de ellos. En este sistema encontramos dos tipos: las listas no superpuestas y los nodos proximales. En ambos casos, almacenan las apariciones de los términos en una estructura jerárquica independiente sobre el mismo texto. Su implementación se basa en la búsqueda de los componentes que concuerden con la cadena determinada, y a continuación se valora cuál de los estos satisface la estructura de la consulta realizada.

3.3.2 Recuperación de información en la red

El nacimiento y desarrollo exponencial de Internet ha traspasado todos los límites inimaginables de la comunicación. Dentro de este ámbito se está generando una gran cantidad de información y que en muchas ocasiones puede ser de gran interés para las empresas a la hora de conocer el entorno.

Dichos sistemas usan métodos similares a los sistemas de recuperación de información tradicionales y debido a esto encontramos algunos inconvenientes a la hora de recuperar la información. Estos problemas son debido principalmente a que el entorno de trabajo no es el mismo, y los datos se encuentran almacenados con distintas características. Además, en Internet surgen nuevos problemas como son el *spam*[1], los *cloakings*[2]...o los vinculados con el enorme tamaño de los índices que pueden llegar a generar grandes cantidades de espacio, así como dificultar la gestión adecuada de los documentos.

La creación de herramientas que faciliten la localización y acceso a esta información, obliga a adoptar dos aproximaciones clásicas de los modelos de recuperación de información automatizada, y que actualmente se utilizan en Internet. Estos son:

- La creación de bases de datos con índices invertidos, mediante la utilización de *robots*, que se encargan de recorrer Internet.

- La creación de listados, catálogos e índices en los que se agrupan contenidos con un tema en común.

Además de las innumerables ventajas de estos sistemas, también existen algunas limitaciones o desventajas procedidas de la organización hipertextual de la Web y de la variación existente en estas.

- Los resultados obtenidos pueden estar solapados por las diferentes interfaces de consulta.

- La cobertura de estos sistemas no es absoluta.

- El usuario tiene la tarea de considerar si los resultados son fiables y de confianza.

1. Información no deseada, no solicitada o desconocida, generalmente recibida o enviada de forma masiva.
2. Técnica por la cual se visualiza contenido diferente al original, con la intención de manipular lo que se indexa.

- Hay diferencias entre el contenido que se almacena y el real.

- En algunos casos, las bases de conocimiento no son actualizadas automáticamente, por lo que puede que las páginas estén desactualizadas.

Asimismo, las problemáticas pueden ser abarcadas desde dos puntos de vista: el humano y el computacional. El primer punto hace referencia al estudio del comportamiento y de las necesidades del usuario, y el segundo se basa en la calidad de los algoritmos y la estructura de datos.

Si nos centramos en el punto de vista del humano, encontramos diversos problemas como:

Uno de ellos es la determinación de la consulta de información por parte del usuario. Si la consulta es demasiado corta o muy general, el sistema recuperará demasiados documentos, limitados en este caso por la cantidad máxima que proporcione cada uno de los servicios de búsqueda. Sin embargo, si la consulta es muy especifica, puede ocasionar que estos no devuelvan ningún resultado. Ambos paradigmas no son validos para el usuario, en el primero el usuario siente una sensación de agobio por la gran cantidad de resultados obtenidos, y en el segundo, una sensación de frustración por no obtener ningún resultado. Una consulta general puede ser refinada hasta obtener una cantidad aceptable de resultados, sin embargo, en el caso de las muy especificas no es posible. Este problema se ve agravado debido al tamaño de Internet, y muchas de las consultas se suponen generales, produciendo una recuperación descomunal de documentos.

La popularidad de Internet ha provocado un incremento del número de usuarios y en su mayor medida con falta de entrenamiento en las búsquedas de información, lo que provoca que los resultados obtenidos en su mayor medida, no sean muy relevantes para estos, limitándose a consultar únicamente los documentos presentados en la primera página. El objetivo de que un usuario acceda al sistema para obtener información, no es con ningún propósito recreativo sino por una necesidad de encontrar y recuperar su necesidad de información. Algunos sistemas, presentan la posibilidad de introducir filtros adicionales que permitan refinar más los resultados obtenidos, aunque a pesar de la existencia de estos filtros avanzados, los usuarios no los suelen utilizar, y algunas de las razones pueden ser:

- Una cualidad que describe al ser humano (y no es muy positiva) es la facilidad que tiene de escoger el camino más sencillo. Debido a esto, en la mayoría de los casos el usuario se comporta de forma vaga, introduciendo una consulta en la caja del buscador y ejecutándola, además de que en algunas ocasiones no ven más allá de los primeros resultados y con ello no reformulan la consulta. Con esto, posiblemente lo único que ocurra, es la obtención de un montón de resultados por el sistema, en los cuales muchos de ellos no contendrán información relevante para el usuario.

- En ocasiones los usuarios pueden saber que quieren buscar, pero carecen del conocimiento o preparación adecuados para expresar su necesidad de información. Una palabra puede tener diversos significados y puede ser expresada por diferentes palabras. Por ejemplo, la palabra *coche* puede estar expresada por *vehículo*, *automóvil*, *auto*...

- Por último, y uno de los puntos más importantes es el desconocimiento de los operadores lógicos. En diversos estudios se ha demostrado que los usuarios no expertos tienen dificultades a la hora de utilizar los operadores proporcionados por los sistemas.

En cambio, si nos centramos en el punto de vista tecnológico, tenemos los siguientes problemas:

Existe un problema que afecta directamente a los sistemas de recuperación de información en la Web basados en interfaces, pero este escapa de su alcance, y es que, los motores de búsqueda de las interfaces de consulta no identifican relación semántica en los documentos. Trabajan con cantidades desorbitadas de información desestructurada pero no contemplan su descripción. Esto ocurre porque el trabajo de indexación recae sobre robots, provocando errores en el tratamiento de la información del código semántico, ocasionado por malas interpretaciones. Este problema no es fácil de resolver, la utilización de metadatos en las páginas Web se hace necesaria para solventarlo, pero continua siendo una tarea que no se lleva a cabo en muchas ocasiones. Si además añadimos que no hay estandarización en el empleo de estos, que se utilizan de forma inadecuada provocando que los motores no siempre los reconozcan y no los utilizan para valorar sus resultados; hace que el problema se vea agravado aún más.

En la siguiente tabla se pueden ver algunas de las principales diferencias del sistema en ambos ámbitos:

	RI Tradicional	RI en la Red
DIVERSIDAD	Homogénea	Muy diversos
ACCESIBILIDAD	Accesible	Parcialmente
FORMATO	Texto	HTML
RESULTADOS	Dependiente	Altísimo
VOLUMEN	Grande	Gigante
CALIDAD	Limpia	Bastante ruido
ACTUALIZACIONES	Poco frecuente	Constante
TÉCNICAS	Basadas en contenidos	Basada en enlaces

Figura 10. Tabla comparativa RI tradicional vs en la red

Uno de los puntos clave en la utilización de estas herramientas, es conocer la calidad de información disponible. En Internet se habla de cualquier tema, donde todo el mundo puede dejar su opinión o perspectivas, sin embargo esto está generando grandes cantidades de información de mala calidad (ruido) que dificultan la localización de información verdaderamente relevante para el usuario. Una manera de abordar esa situación es mediante el acotamiento de la consulta por parte del usuario, ya sea utilizando filtros geográficos, de idioma, por temática o bien con la elección de los diferentes tipos de fuentes disponibles. Sin embargo, aunque los usuarios no tengan muy claro estos parámetros de acotamiento para reducir el ruido, siguen siendo las herramientas por excelencia para captar la información de la red.

Actualmente en Internet se puede buscar información desde muchas fuentes, siendo los principales, los motores de búsqueda, no obstante, debido a la gran cantidad de información y opiniones que subyace de las redes sociales y las aplicaciones de intercambio de contenidos se hace necesario contemplarlos en el proyecto.

3.4 Fuentes de información

A continuación se verá una visión global de los sistemas utilizados en el proyecto para obtener información de Internet, así como, las cualidades que pueden ser de utilidad para la escucha activa en las empresas.

3.4.1 La Web: Motores de búsqueda

Los motores de búsqueda nacen por la necesidad de gestionar las unidades de información en los sistemas de recuperación de información tradicionales. Si miramos atrás en el tiempo, la primera generación

de buscadores data de 1994, cuando Salton y otros investigadores comenzaron a estudiar las estadísticas de las palabras de los documentos, cuyo modelo fue ejemplificado por Altavista.

En 1997, Direct Hit propuso la idea de utilizar los *clicks* como nueva característica de relevancia, y a su vez, Marchiori y otros propusieron también la utilización de los enlaces y el texto de estos como otras medidas de relevancia. En 1998 nacieron dos algoritmos muy conocidos actualmente, el PageRank y los Hits, utilizados por Google.

Más tarde, alrededor del 2003, con la explosión de la Web 2.0 surgieron una serie de buscadores que ofrecían servicios debido a la posibilidad del análisis masivo de la Web. Así pues, comenzaron a aparecer los correctores ortográficos en tiempo real, las sugerencias de las consultas…

Actualmente estos motores se encuentran alojados en servidores Web proporcionando servicios de búsqueda al usuario mediante el acceso a su página Web o a través de sus servicios de consulta por código. Estos motores están basados en la arquitectura cliente-servidor y están compuestos por cuatro módulos:

- Robots: son programas para localizar e indexar los documentos. Parte un listado de direcciones URL y va actualizando la base de datos en función de su visita a la página Web proporcionada. Cada motor de búsqueda implementa su algoritmo, así pues, para Google encontramos el *Googlebot* o para Yahoo! el *Slurp*. Su eficiencia se mide según los siguientes puntos:
    - La potencia del lenguaje de consulta.
    - La cantidad de información almacenada por cada página Web.
    - La eficiencia para descubrir nuevas páginas Web y mantener la lista conocida.
- Base de datos: donde se gestiona y almacena la información extraída de las páginas Web. La lista de información almacenada varía según cada sistema de recuperación, pero suelen tener en común el titulo, la URL, un texto descriptivo o las palabras claves.
- Búsqueda: algoritmos de localización y recuperación de documentos, que gestionan las peticiones de consulta por parte de los usuarios.
- Interfaz de consulta: donde el usuario podrá realizar una consulta al sistema y visualizar los resultados más relevantes devueltos por el sistema de búsqueda.

3.4.2 La Web: Directorios

Este sistema se centra en la generación de categorías por temática para facilitar las búsquedas a un usuario. Las bases de conocimientos de estos sistemas se diseñan para que un mismo documento pueda estar agrupado en un mismo tema, siguiendo un orden jerárquico en forma de árbol, de los más generales a los más concretos. Las ramas principales de este árbol pueden ser vistas como los temas principales, y sus ramificaciones serían las subcategorías. El administrador del sistema es el encargado de ir formando la base de datos mediante la inserción o eliminación de una URL a la jerarquía.

Entre las virtudes de estos sistemas cabe destacar la posibilidad de conocer a priori cuales son las páginas Web con más relevancia que pertenecen a un mismo tema, además de que este tipo de sistemas ofrecen una buena calidad de indexación de estas. Por contra, están penalizados con una mayor lentitud frente a los buscadores, además debido a que la base de conocimiento es mantenida de forma manual, cabe la

posibilidad de que una página Web no se encuentre disponible porque ya no exista.

Actualmente, estos servicios han ido evolucionando e incorporando nuevas prestaciones, llegando ha convertirse en puertas de acceso al mundo de Internet, denominados portales[1] Web.

Un ejemplo de este tipo de sistema, y que además se ha utilizado su servicio de consulta para el sistema de recuperación de información desarrollado en el proyecto es Yahoo!. Aunque también existen otros como, *Web Katalog o Portal-SEO*.

3.4.3 Redes sociales

Actualmente las redes sociales son un aspecto importante de cada persona en el día a día. Un alto porcentaje del tiempo invertido en la navegación por Internet, se utiliza para ingresar a alguna de las redes sociales, llegando a tal nivel, que siete de las redes sociales más conocidas se encuentran entre los cincuenta primeros puestos con mayores visitas de todo el mundo. Los usuarios crean perfiles a través de los cuales administran y generan nueva información, pero hay que destacar, que el acceso a esta información esta ligada al grado de privacidad que dichos usuarios establezcan para la misma, por lo que se convierte en una de las principales barreras de los sistemas de recuperación de información en estos entornos.

Existen dos tipos de redes sociales desde el punto de vista empresarial, las enfocadas a los consumidores y las enfocadas a las empresas.

En las primeras, redes como Facebook, tienen como objetivo la interconexión de usuarios con una afinidad común. Dentro de esta red, los usuarios pueden completar su perfil mediante la inserción de sus datos básicos, vídeos, fotografías...e ir realizando conexiones comunes en su red de contactos, de forma que puedan compartir información mutuamente. La facilidad con la que el usuario puede publicar información, noticias, contenido multimedia, opiniones...hace que se convierta en uno de los canales más populares.

No obstante, en estas redes las empresas tienen la posibilidad de generar páginas en las se pueden promocionar, ofreciendo al usuario información actualizada, noticias, sus productos...y de esta forma interactuar con los usuarios, dado que estos pueden ir dejando comentarios y opiniones respecto a los servicios ofrecidos.

En las segundas, al ser unas redes sociales específicas para las empresas, el enfoque es algo distinto, pero que desde una visión de la escucha activa son igual de importantes. En este caso, este tipo de redes esta orientada al *marketing online* [B.31] de las empresas. Las sugerencias, el conocer mejor a la competencia y cuales son sus servicios, el mantener el contacto con expertos del sector, el tener una estrecha relación con los clientes…son algunos de los puntos clave de este tipo de redes sociales, y que pueden ser de gran ayuda a las empresas para darse a conocer y mantener una relación activa con los usuarios.

Las posibilidades de comunicación que ofrece las redes sociales, tanto para los usuarios como para las empresas, están generando constantemente grandes volúmenes de información, donde se pueden encontrar desde simples anotaciones que no son de gran importancia, hasta verdaderas opiniones que son de gran utilidad para las empresas. Es por ello, que se hace interesante la captación de toda esta información en la escucha activa, y así poder ayudar a las empresas, por ejemplo, en el acercamiento al usuario, en la mejora de sus productos o en su estrategia de *marketing online.*

- Es un sitio Web donde se ofrece al usuario una serie de servicios que tienen como objetivo satisfacer las necesidades de acceso a una serie de recursos.

3.4.4 Microblogging

El funcionamiento de este tipo de fuente se basa en la publicación de mensajes cortos (máximo de 140 caracteres), que son publicados en la página del perfil del usuario (timeline[1]). Además, también serán enviados a todos aquellos que se han suscrito a la opción de seguir a un usuario concreto (followers[2]). Debido al poco espacio que ofrece para comunicarse, los mensajes suelen ser mucho más directos y con un impacto mayor.

Algunas de las principales características de estas plataformas es la interacción entre los usuarios y su facilidad de hacerlo. La interacción con los usuarios es uno de los aspectos más importantes, un usuario puede publicar un mensaje en su perfil, pero también puede replicar uno de otro usuario que le parezca interesante. De esta forma, el mensaje es propagado al resto de contactos del usuario. Al mismo tiempo, puede que otros usuarios lo encuentren interesante y lo vuelvan a republicar, llegando de esta manera a un mayor número de usuarios. Se puede apreciar, que un mensaje con impacto y que sea de interés para los usuarios puede llegar a alcanzar una propagación exponencial por la red.

A diferencia de otras plataformas, los microblogging permiten crear una comunicación entre las empresas y los usuarios, la cual hay que utilizar para crear una relación de acercamiento a estos. Esta comunicación se caracteriza por ser bidireccional, así pues no solo hay que transmitir información a los usuarios, sino que lo atractivo de esta, es la posibilidad de mantener una actitud de escucha activa con los usuarios.

Lo más importante desde el punto de vista de la escucha activa, es que en este tipo de herramientas la información sucede prácticamente en tiempo real, ofreciendo así contenidos actualizados sobre un tema o evento que puede ser de interés en un momento dado para las empresas o los usuarios.

Si se obtiene toda esta información y es utilizada de forma adecuada, puede brindar a las empresas abundantes beneficios en muchos aspectos. Cuando esta información es utilizada para mejorar la comunicación con los usuarios, entendiéndoles, respondiéndoles, escuchándoles…en definitiva, cercando más las distancias entre empresa y usuario, estas pueden llegar a crear un entorno de comunicación, donde los usuarios aportan nuevos contenidos que pueden llegar a ser de un alto grado de interés para estas. Además, debido a la capacidad de expansión de los mensajes, las empresas pueden conseguir una mayor propagación de sus contenidos, así como captar nueva información de usuarios que generen contenido relevante.

3.4.5 Multimedia

Otra fuente de información y no menos importante, son las herramientas Web 2.0 de multimedia, cuyo objetivo principal es la compartición de contenidos como imágenes, vídeos, audio...en comunidades de usuarios o empresas que los publican gratuitamente en Internet. En este tipo de herramientas, no existe un perfil definido, ya que son utilizados de igual manera por los usuarios como por las empresas. Los servicios que ofrece se centran en la diversión, el fomento, la publicitación gratuita mediante presentaciones o vídeos, el marketing online…entre otros.

Los principales beneficios que puede aportar la escucha activa en este tipo de fuentes son el impacto y la visibilidad de la empresa en estos medios. A través de este canal, las empresas pueden mantener una comunicación con los usuarios de una forma más llamativa, visualmente hablando.

1. Página donde se visualizan todos los tweets asociados con un usuario y sus relaciones.
2. Son todos los usuarios que están conectados o relacionados con un usuario en concreto.

3.5 Sistemas de escucha activa

3.5.1 La importancia de la escucha activa en las empresas

El ser humano ha ido evolucionando a lo largo de los años con un único objetivo, la supervivencia, y para ello ha tenido que someterse a una gran variedad de condiciones. En cambio, está diseñado para eludir las complejidades que se le presentan, y dar una respuesta lo más rápida y breve posible. Sin embargo, en el mundo empresarial actual se necesita de una mejor precisión en las respuestas para la toma de decisiones.

Gracias a la aparición de Internet y a los servicios de búsqueda, las empresas tienen la posibilidad de obtener grandes cantidades de información que puede ayudarles a entender lo que esta sucediendo en su entorno. El problema, es que existen varios factores humanos que limitan este proceso. El ser humano no está preparado para ser abordado por tan gran cantidad de información y además es frágil a la hora de tomar decisiones importantes. Si a esto, le sumamos las exigencias empresariales a las que tiene que hacer frente, se hace necesaria la utilización de sistemas que ayuden a la empresa a captar y organizar toda esa información para facilitarle la toma de decisiones.

Generalmente las empresas son conocedoras del impacto que tienen en el mercado, pero ¿son conocedoras del impacto en Internet?, ¿saben lo que se dice de ellos?, ¿qué alcance tiene?... Es importante que las empresas sepan responder a estas preguntas, y para ello necesitan la capacidad de obtener la información con la mayor exactitud posible de los medios *online* en los que puede que este presente, y así aporta un valor añadido no solo fundamentado en la prevención, sino también en convertirlo en un valor estratégico.

Debido a la aparición de las nuevas tecnologías de la Web 2.0, las empresas pueden aprovechar más que nunca esta información. Las palabras recoger, analizar, comprender, escuchar y actuar han adquirido un significado totalmente distinto al conocido. De hecho, la aparición de términos como la sabiduría colectiva, software colaborativo...o herramientas basadas en la Web 2.0 como las redes sociales, *blogs*, *microblogs*...forman el cambio paradigmático en que las empresas toman actualmente sus decisiones. ¿Pero cómo se aborda? Estar pendientes de todas estas plataformas Web 2.0 es imposible, es por ello que se precisa en este tipo de aplicaciones de un marco para recuperar toda la información y qué se pueda entender que tipo de información es útil, deseable y asequible para las empresas.

Conocer, analizar, actuar y retroalimentar son términos que implica básicamente dos tareas de alto nivel: la primera es la generación de soluciones, que implica la obtención de información, la detección de problemas y las hipótesis de resolución, y la segunda es la evaluación del impacto y las consecuencias por las acciones tomadas.

A la hora de generar soluciones a un problema, las personas tienden a buscar información que confirme su hipótesis y mantener tales creencias a pesar que puede que existan algunas contradicciones. Se tiende a ver patrones donde nos los hay y a dejarse influenciar por soluciones que se presentan de forma atractiva. La captación de más información por otras vías y la evaluación de ésta, son factores que pueden ayudarles en la toma de decisiones, ofreciéndoles la posibilidad de entender, analizar y comparar la información.

Por todo ello, muchas de las empresas están utilizando sistemas que les ayuden a captar información, para ayudarles a aumentar su conocimiento y así poder mantener una actitud de escucha activa con los usuarios.

El uso de estas aplicaciones es muy diverso, entre los cuales se destaca la examinación de mercados, la atención a los usuarios y la gestión del conocimiento. Obviamente, el determinado uso que se le de a estas aplicaciones afectará al modo en el que se debería de evaluar el éxito, además de que existen determinados indicadores que solo pueden ser evaluados una vez terminado el proceso.

Aunque los indicadores pueden ser diferentes, un indicador clave es el conocimiento, es decir, si la aplicación a generado suficiente conocimiento y de buena calidad como para poder evaluar y actuar en función de esta.

3.5.2 Algunas de las aplicaciones

Desde hace varios años se esta viviendo un crecimiento continuo del marketing en redes sociales y junto a este ha surgido un mercado de software de gestión y monitorización de estos entornos. Este crecimiento acelerado ha hecho que exista un mercado de servicios, proveedores y herramientas para medir o gestionar el social media. Se presenta a continuación un resumen a través de ejemplos de algunas de las aplicaciones de escucha activa que se pueden encontrar en Internet.

BrandChats [E.23], es una herramienta Web 2.0 que trabaja analizando la información existente en la red social Facebook, en el sistema de microblogging de Twitter, y en las aplicaciones de intercambio de contenidos como YouTube y Flickr. La captación de información se puede configurar para obtener información tanto de la empresa como de su competencia, permitiendo filtrar, analizar y editar los resultados obtenidos.

Algunas de sus características más importantes son (*):

- Informar de cómo hablan de la empresa, identificando quién lo ha dicho, su poder de conexión y los influenciadores *online* que tiene.
- Identificación de la fuente de origen (*blogs*, microblogs, video...). Geolocalización de donde es nombrada la empresa y reconocimiento del idioma.
- Análisis de opinión, de conceptos y valores asociados a la empresa. Generación de ideas para mejorar la reputación de la empresa. Control de imagen corporativa.
- Análisis del posicionamiento (SEO, en inglés *Search Engine Optimization*).
- Diseño de las estrategias para la comunicación.
- Localizar a las personas influyentes.

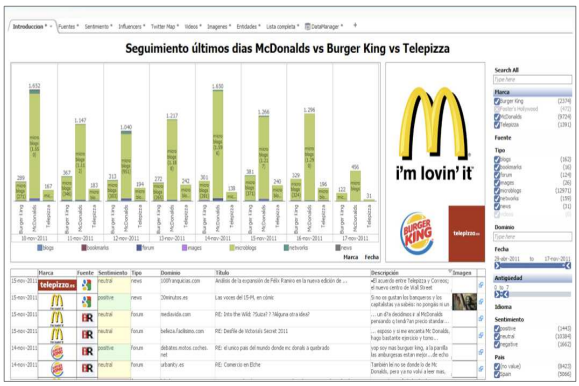

Figura 11. BrandChats

(*) Información obtenida de la página Web

Radian6 [E.24], utiliza todos los medios sociales para obtener información relevante para la empresa. Esta aplicación fue lanzada en el Dreamforce 2011, evento de referencia en el ámbito de la computación en la nube, que se llevó a cabo del 30 de agosto al 2 de septiembre, en San Francisco.

A partir de una configuración previa, donde se seleccionaran las palabras claves a buscar así como los diferentes filtros, el sistema comenzará a gestionar la red en busca de información importante para el usuario.

Entre sus características principales encontramos (*):

- Escuchar, medir, participar y descubrir la información que fluye por las redes sociales.
- Encontrar información que nombre a la empresa, además de proporcionarle sugerencias para optimizar los resultados.
- Panel resumido donde se expone toda la información.
- Permite una comunicación directa con los usuarios a través de la aplicación.
- Rastrea más de 150 millones de fuentes. Si esto no fuera suficiente, permite agregar nuevas fuentes de datos a la indexación.
- Ofrece un histórico de datos de aproximadamente de un mes.
- Personalización en el análisis de opinión.
- Análisis en tiempo real o selección de fechas.
- Inteligencia competitiva.

Figura 12. Radian 6

(*) Información obtenida de la página Web

Socialmention [E.25], se comporta igual como un buscador. Pero en este caso las búsquedas se centran en los comentarios de los usuarios en las redes sociales, microblogging, *blogs* y socialbookmarks. Utiliza cuatro variables para medir las características de las consultas presentadas por el usuario (*).

- Fuerza: es la probabilidad de que la empresa se esté nombrando en las redes sociales.
- Sentimiento: promedio de frases positivas y negativas.
- Pasión: probabilidad de que las personas que han mencionado tu empresa, vuelvan a hacerlo.
- Alcance: media del rango de influencia. Es el número de autores únicos que hacen referencia a la empresa.

Y entre sus características podemos encontrar (* Información obtenida de la página Web):

- Aunque dispone de una extensa variedad de páginas Web donde encontrar la información, su nicho se centra en las redes sociales.
- Respuesta en tiempo real.
- Filtrado de datos mucho más sencillo, pero también limitado.
- Histórico de datos aproximado de dos meses.
- Análisis de opinión automático y métricas propias de la aplicación, como hemos visto anteriormente.
- Es gratuita.

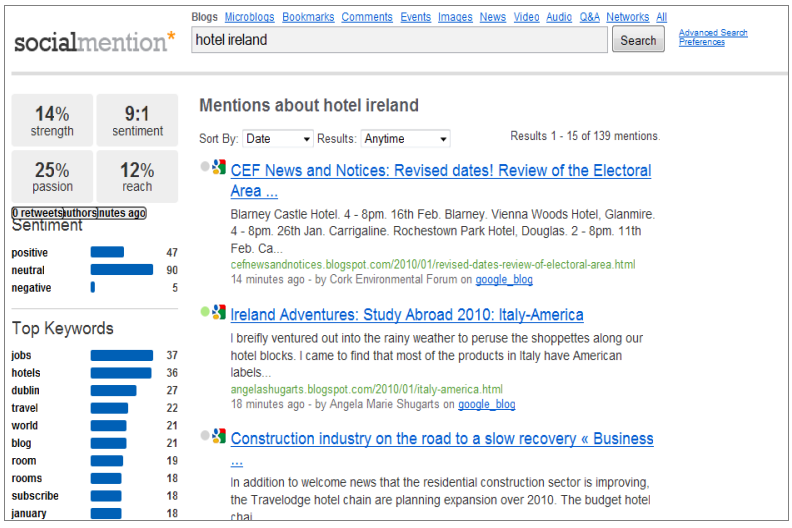

Figura 13. SocialMention

(*) Información obtenida de la página Web

4. VISIÓN GENERAL DE COSMOS

4.1- Descripción

La plataforma de escucha activa fue bautizada no hace mucho como Cosmos [E.26] [B.32], y es una herramienta desarrollada por Autoritas Consulting cuyo objetivo principal es gestionar todas las necesidades que una institución necesita obtener de Internet [B.33], ayudándola a ordenar todo el caos que existe en la red, ofreciendo técnicas para los usuarios o grupos de trabajo que desean profesionalizar las oportunidades que se producen en este ámbito. Sus tres principales objetivos son: Escuchar, Analizar y Actuar.

El rastreo de Cosmos en Internet abarca cuatro puntos principales:

- Reputación online: búsquedas de contenido concretos (geolocalizados o no), llamadas también búsquedas textuales. Se centra en la exploración de palabras o fragmentos de textos en Internet relacionadas con la marca búsqueda. Los resultados son asociados a diferentes grupos:
    - *Organización*: La cual abarca conceptos relacionados con la organización, marca, productos, personas y campañas.
    - *Competencia*: Marcas competidoras directamente ligadas con la buscada.
    - *Mi acción*. Búsquedas que permite tener las medidas del propio impacto social.
- Actividad: los datos obtenidos provienen de diferentes fuentes que se han de identificar [B.34][B.35]. Mediante un exhaustivo análisis y *tracking* se obtienen las acciones relacionadas por la marca.
- Target: basado en la captura de toda la información activa que se produce en determinadas áreas geográficas [B.36], en un determinado idioma [B.37] y/o listas de clientes.
- Inspiración: las búsquedas están centradas en determinadas *URLs*, en la Web en general, en redes sociales, en herramientas de intercambio de contenidos…recuperando todo lo que se está diciendo.

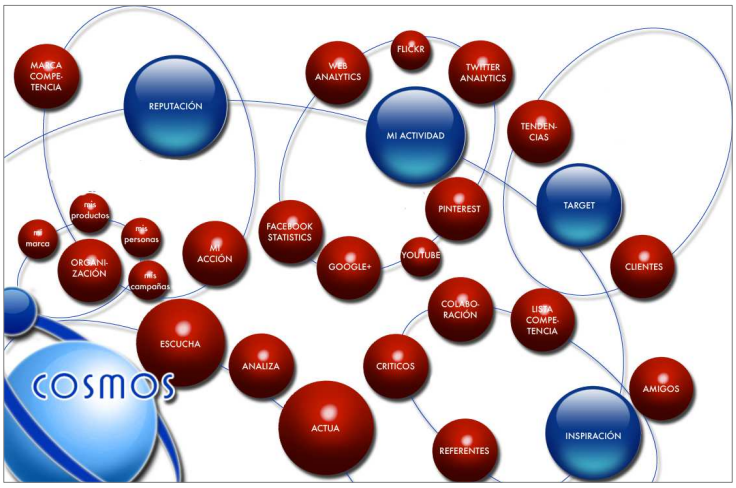

Figura 14. Descripción de Cosmos

4.2. Estrategia de comunicación

Internet y los medios sociales son herramientas que permiten a los usuarios comunicarse y relacionarse con otros usuarios, además de aportar contenidos y nueva información. Por lo que, actualmente el control de las marcas está en manos de los usuarios.

Actualmente, con las tecnologías que la Web 2.0 ofrece, se hace necesario escuchar lo que los usuarios opinan de una empresa, pero además, no solo es escuchar, sino atender y conversar con los usuarios con el objetivo de obtener nuevas experiencias. Así pues, para una comunicación efectiva, se ha de haber definido una buena estrategia de comunicación, estableciendo que es la empresa y que servicios ofrece.

Si no esta bien definida la estrategia de comunicación, puede resultar muy difícil alcanzar los objetivos planteados, por ello, la herramienta Cosmos, facilita el entendimiento a partir de una situación inicial y permite configurar y reconfigurar las estrategias a seguir en Internet en función de la realidad existente. Para conseguir esto, Cosmos se basa en los siguientes puntos:

- Definición de objetivos: es el punto principal de toda estrategia de comunicación. Deben de ser siempre bien planificados, tenerlos bien definidos y saber cómo se van a conseguir.

- Definición del target: es necesario conocer cual es el perfil y el comportamiento de los usuarios para la elección adecuada de los canales a nuestras necesidades [B.38].

    - *Activistas*: usuarios que adquieren y generan información o forman vínculos a sus intereses.

    - *Clientes*: usuarios que se relacionan directamente con la empresa, y pueden crear enlaces de información directa.

    - *Profesionales*: usuarios que trabajan en el mismo sector que la empresa, y que engloban a la competencia.

    - *Inversores*: usuarios del sector financiero que consumen información económica de la empresa.

    - *Periodistas*: usuarios profesionales que informan de las últimas noticias.

- Identificar que canales sociales son de mayor impacto para la estrategia.

- Generación de tareas de comunicación en los canales sociales participativos.

Cosmos se comporta como un *wizard* estratégico que a partir de los objetivos establecidos permite generar unas estrategias a seguir.

4.3. Internet Scorecard

El concepto de cuadro de mando integral [B.39] (BSC, del inglés *Balanced Scorecard*) fue introducido en el número de Enero de 1992 de la revista de *Harvard Business Review*.

Robert Kaplan y David Norton, los autores, definían las cuatro perspectivas posibles del CMI, en las que una organización, en su componente financiera, de clientes, de actividades internas o de conocimiento, podía desenvolverse, siendo este, un sistema con el que evaluar las funciones de una institución según su estrategia.

Figura 15. CMI

- Perspectiva financiera: se centra en la parte económica de la empresa y muestran su pasado.
- Perspectiva clientes: mide los enlaces con los clientes y los pronósticos de los negocios. Se considera que esta perspectiva es la más importante debido a que va ligada directamente con los consumidores de la empresa, por tanto se han de cumplir todas las expectativas de los clientes.
- Perspectiva de procesos internos: identifica los procesos críticos en los que la empresa ha de deslumbrar cara al cliente.
- Perspectiva de formación y crecimiento: mide las necesidades de la empresa para mantenerse y crecer a un largo plazo de tiempo.

Al ingresar al nuevo milenio, Internet ha modificado los habituales modos de hacer negocio, viéndose transformadas las conductas de los clientes y los empresarios. El uso de este medio ha provocado una modificación en los métodos de marketing de las empresas, así como en los servicios de distribución.

En 1992, no se había integrado Internet en las instituciones, por lo que después del gran éxito de este se debe de contemplar este medio en el CMI de las organizaciones.

La herramienta Cosmos incorpora una nueva perspectiva orientada a Internet (*Internet Scorecard*) al cuadro de mandos integral CMI de las organizaciones, facilitando la medición de los objetivos a alcanzar y su grado de cumplimiento (KPI, del inglés *Key Performance Indicators*). Para conseguirlo, a partir de los ámbitos de impacto de Internet posibilita un cuadro de mando adecuado para cada una de las áreas de la organización.

4.4 Inteligencia social

El término inteligencia social tampoco es un término nuevo, sus inicios datan de los años noventa, donde se estableció un significado reconocido por el libro Inteligencia Emocional [B.40] (del inglés, *Emotional Intelligence*).

Este término puede resumirse como las posibilidades de comprender y controlar los sentimientos e interpretaciones que subyacen de otras personas. Pero no solo en las relaciones con los demás, sino también en los sentimientos de uno mismo.

Años más tarde, el mismo autor publicó el libro Inteligencia Social [B.41] (del inglés, *Social Intelligence*), con el cual profundizó en las relaciones personales.

Con las nuevas tecnologías de la Web 2.0 y la revolución de las redes colaborativas, el término de inteligencia social impuesto por Daniel Goleman, ha adquirido un significado más fuerte, donde los usuarios comparten información en comunidades como conjunto de una inteligencia colectiva global.

La herramienta Cosmos, exprime toda la información recuperada, extrayendo relaciones de toda la información y permitiéndole a la organización navegar entre lo que se encuentre relacionado. De esta forma, la herramienta aprende de los usuarios para la generación de metaontologías [B.42].

4.5 Influenciadores

Los influenciadores se consideran a aquellos usuarios que son generadores de información [B.43] especializados en un tema concreto y que además son consumidores de esta. El que un usuario tenga más seguidores o publique mayor cantidad de contenidos no necesariamente quiere decir que sea un usuario influyente. Además, una cualidad que los identifica es la gran capacidad de participación que ofrecen a los usuarios, compartiendo experiencias e ideas. Para ello, se utilizan técnicas de reconocimiento de entidades con nombre (NER, en inglés *Named-entity Recognition*) [B.44] para identificar las personas influyentes en la escucha activa.

La herramienta Cosmos, identifica a los influenciadores basándose en las seis mediciones para la identificación de la influencia publicadas en el *Harvard Business* por David Armano [B.45]:

- Alcance: mide los medios en los que se divulga el mensaje. Un usuario puede hace llegar su mensaje por diferentes canales de información que presenta la Web 2.0.

- Experiencia: mide cuanto es de experto un usuario. Se basa en el análisis del conocimiento por parte del usuario, viendo el conjunto participativo que aporta en las redes colaborativas.

- Relevancia: mide el rendimiento de un usuario dentro de una comunidad colaborativa.

- Credibilidad: la información trasmitida ha de estar contrastada, proporcionando seguridad a los usuarios de obtener una información certera.

- Confianza: vinculo hacia el usuario creado por la credibilidad de la información y la familiaridad de estos.

- Proximidad: mida las relaciones entre los usuarios y como están conectadas.

5. TECNOLOGÍA UTILIZADA

En el siguiente capítulo se detallan cada una de las tecnologías utilizadas para desarrollar el proyecto así como sus lenguajes de programación. En cada uno de los apartados se muestran sus características principales y una justificación de porque se ha utilizado esa tecnología.

5.1 Gestor de contenidos

Todo el sistema ha sido desarrollado utilizando el gestor de contenidos Liferay [E.27] [B.46]. Este gestor se caracteriza por ser de código abierto y estar escrito en Java. Los inicios de este gestor de contenidos datan del año 2000, y surge por la necesidad de dar soluciones corporativas a las organizaciones, ayudándolas a desarrollar soluciones empresariales con resultados inmediatos y con un valor a largo plazo. A continuación se presentan algunas de las características de porque se ha utilizado este gesto de contenidos en el proyecto.

Entre las posibilidades que ofrece, es la creación de portales Web, donde de forma sencilla se pueden construir aplicaciones y sitios Web con altas prestaciones. Estos portales generalmente se construyen a partir de un conjunto de portlets o *gadgets* desarrollados para las páginas que componen el portal.

Todo portal Web desarrollado con el gestor de contenidos, está compuesto por cuatro módulos:

- Tema: apariencia del portal, que puede ser común o no a todo el sitio.
- Portlets: es el punto más importante del portal. Un portlet puede ser definido como una pequeña aplicación de funcionalidad limitada, que en conjunto con otros confortan un sistema más complejo. Por su naturaleza similar a los servlets, estos son manejados y controlados por el contenedor, producen contenido dinámico e interactúan con el cliente Web a través de la utilización de peticiones request – response.
- Páginas: todo portal está compuesto por páginas enlazadas entre sí. Gracias al gestor de contenidos, se pueden generar nuevas páginas sin tener que escribir nuevo código, con la posibilidad de reutilizar los portles existentes.
- Sistemas de navegación: conjunto de menús que facilitan la navegación y la agrupación del contenido del portal.

La creación de páginas personalizadas, así como la separación entre usuarios anónimos y autentificados se hace de forma sencilla a partir del panel de configuración. Se pueden crear páginas, donde unos usuarios ven un contenido distinto al resto de usuarios. Así pues, la generación de comunidades se reduce a la asignación de permisos a los usuarios que se desee que trabajen de forma colaborativa en la comunidad.

El sistema ofrece la posibilidad de mostrar el portal en diferentes idiomas. Mediante la edición de unos ficheros de diccionarios ha sido posible modificar la información expuesta en el portal a diferentes idiomas.

Como el sistema está basado en los principios de las aplicaciones Web, este puede ser visto en diferentes plataformas, independiente del sistema operativo o dispositivo utilizado.

Centrándonos en el aspecto técnico, Liferay es compatible con los sistemas operativos Linux y Windows. Ofrece la posibilidad de utilizar tres tipos de contenedores de *servlets*: Jetty, Tomcat y Resin. En nuestro caso, hemos escogido Tomcat, debido a que era un sistema ya conocido, además de que en el portal de Liferay proporciona en su descarga el contenedor totalmente configurado. A su vez, Liferay es compatible con muchos tipos de bases de datos: MySQL, Postgres, Oracle...

Liferay proporciona un SDK compatible para Eclipse o Netbeans, con el que facilita su integración y configuración del portal, además de presentar características que facilitan el desarrollo de los portles.

5.2 Diseño

Debido a que el sistema esta orientado a ser una aplicación Web, es importante conocer los lenguajes de programación utilizados para la creación de las páginas Web. Se ha decantado por HTML y no XHTML principalmente porque el segundo es mucho más estricto en cuanto a los requisitos establecidos.

5.2.1 HTML

El lenguaje HTML [B.47] (del inglés *Hypertext Markupp Language*) es un estándar para las páginas Web. Ese fue creado por el consorcio internacional W3C (*World Wide Web Consortium*) como lenguaje para el diseño y estructuración de textos para su representación en forma de hipertexto.

En este lenguaje se escribe utilizando pares de etiquetas, donde cada una de estas tienen un uso y significado propio. Algunos de los funcionamientos de estas etiquetas son: tablas, inserción de imágenes, listas, líneas de encabezamiento, formato de textos, enlaces... Cada una de estas etiquetas tienen dos cualidades: los atributos y el contenido.

Los atributos son valores puestos detrás del nombre de la etiqueta separados por un = que denotan una cualidad de esa etiqueta, mientras que el contenido está situado entre las dos etiquetas y es el elemento al que se le aplica la etiqueta y los atributos.

En el siguiente ejemplo se ve la estructura de estos tres elementos:

```
<etiqueta atributo="valor">Contenido</etiqueta>
```

Figura 16. Etiquetado HTML

Además de tener en cuenta las especificaciones propias del lenguaje HTML, es recomendable que se siga unas pautas de accesibilidad Web cuando se desarrollan aplicaciones utilizando este lenguaje. De esta forma, se asegura el acceso a la información por todas las personas, independientemente de si tiene discapacidades físicas o si la tecnología no es la adecuada.

El uso de este lenguaje es sencillo. Se puede desarrollado a través de cualquier editor de texto o bien desde un programa especializado en edición de páginas Web mediante HTML.

Actualmente la última versión se encuentra en la cinco. En ella se especifican dos tipos de sintaxis posibles en un mismo desarrollo: el HTML y el XHTML.

5.2.2 CSS

Las hojas de estilo [B.48] (CSS, del inglés *Cascading Style Sheets*) es un lenguaje por el cual se define el aspecto visual de un documento desarrollado en HTML o XHTML. Su finalidad es la de separar el contenido de la estructura visual, introduciendo para ello mecanismos de representación de la información de estilo permitiendo a los desarrolladores representar múltiples páginas Web al mismo tiempo sin tener que modificar la estructura principal.

La sintaxis de un CSS es muy simple. Su estructura se basa en dos elementos: los selectores y las propiedades. A continuación puede verse cual como es la estructura:

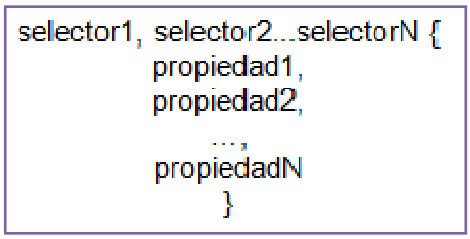

Figura 17. Estructura CSS

Existen dos principales formas de utilización de las hojas de estilo: interna o externa.

Las hojas de estilo internas consisten en la introducción de código de estilo dentro del documento HTML, mediante la utilización de la etiqueta <*style*> justamente dentro del elemento HTML <*head*>. De esta manera se consigue separar el estilo de la estructura de la páginas Web, aunque si se necesita modificar algún aspecto visual, es necesario ir a cada uno de los ficheros de la estructura y cambiar cada uno de los elementos deseados.

Por otra parte, las hojas de estilo externas, son archivos que se almacenan fuera del fichero de la estructura. Mediante la inclusión de la etiqueta <*link*> o la propiedad *import* de HTML se puede incluir cada uno de los ficheros de estilo desarrollados. Esta forma es la manera más potente de desarrollar las hojas de estilo, debido a que separa completamente las reglas visuales del formato estructural de la página Web.

Actualmente se encuentra disponible la versión tres de este lenguaje, cuyas especificaciones añaden nuevas funcionalidades a la anterior versión de manera que se pueda preservar la compatibilidad.

5.3 Información recuperada

Debido a la cantidad de interfaces de recuperación de información utilizadas en el proyecto, y a que cada una de ellas presenta diversos lenguajes en los que los resultados obtenidos son representados, se hace necesario ver una visión general de las características de estos y sus estructuras. Existen varios tipos de representación, de los cuales se van a ver dos por ser los más comunes entre las diversas interfaces.

5.3.1 XML

El lenguaje XML [B.49] (del inglés *Extensible Markup Language*) fue creado por el W3C como adaptación y mejoramiento del lenguaje SGML (del inglés *Standard Generalized Markup Language*) para la ordenación y etiquetado de documentos.

Su nacimiento no es exclusivo para Internet: se plantea como un lenguaje estructurado de información para el intercambio de datos entre dos plataformas. De esta forma, se puede utilizar en bases de datos, editores de texto, hojas de cálculo...

Este lenguaje es un conjunto de módulos que sirven para la estructuración, almacenamiento e intercambio de información. Su verdadera importancia radica en que permite la compatibilidad entre diversos sistemas para el intercambio de información de manera ágil, eficaz y segura.

Entre algunas de sus ventajas cabe destacar la sencillez de entender su estructura y procesarla, la capacidad de generar nuevas etiquetas una vez diseñada y puesta en producción, su componente estándar que posibilita la utilización de cualquier analizador de estos documentos. No obstante, cabe destacar que un

documento XML ha de estar bien formado, es decir, debe de cumplir con las definiciones del formato expuestas del lenguaje, para poder ser analizado correctamente por cualquiera de sus analizadores sintácticos.

5.3.2 JSON

El lenguaje JSON [B.50] (del inglés *JavaScript Object Notation*) es un subconjunto del lenguaje de objetos de JavaScript, cuyo objetivo es la de intercambiar información de forma ligera.

Debido a su facilidad, generalidad y sencillez se ha convertido en el principal sustituto del XML, siendo su principal ventaja la facilidad de desarrollar un analizador para entender este lenguaje.

Su estructura de basa en dos componentes: una recopilación de pares de nombre-valor denominado objetos o registros, y una lista ordenada de valores denominados vectores o arreglos. A continuación se muestra una pequeña porción de una posible estructura:

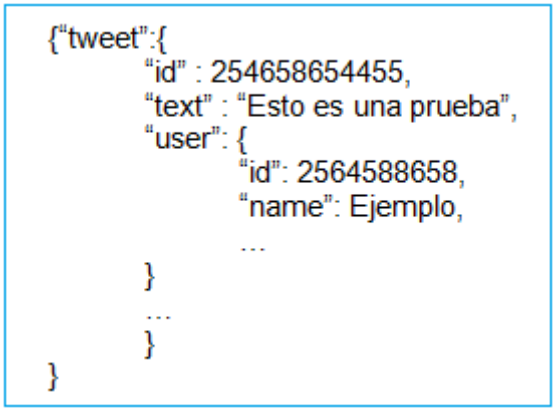

Figura 18. Estructura JSON

La sustitución de JSON frente a XML es algo frecuente, pero normalmente las aplicaciones que lo utilizan suelen brindar la posibilidad de los dos lenguajes.

5.3.3 Comparación de ambos lenguajes

El lenguaje XML tiene un mayor soporte en cuanto a la disposición de más herramientas para su creación y edición tanto para el lado del servidor como para el del cliente.

Los ficheros XML son más extensos y ofrecen mayor dificultad a la hora de analizarlos.

El lenguaje JSON es más compacto por lo que favorece la eficiencia en el tratamiento del fichero.

El lenguaje JSON tiene más analizadores en el lado del servidor.

Ambos lenguajes no tienen mecanismo para grandes objetos binarios.

5.4 Programación

En el siguiente punto se verán las plataformas de desarrollo utilizadas así como los lenguajes de programación usados para el desarrollo Web como para la implementación de las interfaces de consulta.

5.4.1 Entornos de desarrollo

Debido a que el gestor de contenedores Liferay ofrece el SDK de desarrollo para Eclipse [B.51], se ha decantado por este entorno de desarrollo.

Este entorno de desarrollo se caracteriza por ser de código abierto y multiplataforma, además de incluir una gran comunidad de usuarios activa, que da soporte extendiendo continuamente las áreas de aplicación del entorno. Sus inicios se realizaron por la empresa IBM, pero actualmente, es mantenida y desarrollada por una organización sin ánimo de lucro que fomenta el código libre, llamada Fundación Eclipse.

Aunque la última versión de este entorno se encuentra en la *4.2 (Juno),* para el desarrollo de los portlets se ha utilizado la versión anterior (3.7 Indigo), debido principalmente a que el SDK de Liferay es completamente compatible con esta última.

En cambio, para la generación de las diferentes interfaces de consulta a los servicios de búsqueda utilizados en el proyecto, se ha utilizado el entorno de desarrollo NetBeans [B.52].

Este entorno de programación también es libre, sin restricciones de uso, y su desarrollo esta principalmente enfocado al lenguaje de programación Java, aunque existen diferentes módulos de integración para extender al resto de lenguajes.

NetBeans tolera que las aplicaciones se desarrollen desde un conjunto de módulos. Estos módulos son una serie de ficheros Java que puede interactuar con las distintas APIs y un archivo principal que lo identifique. Debido a que los distintos módulos se pueden desarrollar independientemente, se hace de forma sencilla el extenderlos a diferentes desarrollos de software.

Igual que con Eclipse, el proyecto NetBeans es de código abierto y cuenta con una comunidad de usuarios que la mantiene. Nació en el año 2000 como proyecto para mejorar el desarrollo de aplicaciones Java. Actualmente su patrocinador es *Sun MicroSystems*.

NetBeans se encuentra en su versión 7.1 donde su mayor característica es la mejora de soporte para los frameworks como Hibernate y Struts.

5.4.2 Java

El lenguaje de programación Java [B.53] es un lenguaje de alto nivel orientado a objetos, desarrollado en 1995 por James Gosling, con la intención de establecer un lenguaje que pudiera trabajar en las redes de computadores heterogéneas, además de ser autónomo de la plataforma donde se ejecutara.

Actualmente, *Sun Microsystems* se encarga de las especificaciones, el progreso y la evolución del lenguaje, a través de la comunidad *Java Community Process*, de manera que la mayoría todo el lenguaje de programación Java es de software libre.

Algunas de las características del diseño de este lenguaje son:

- Similitud de sintaxis con lenguajes de programación ya conocidos.
- A través del recolector de basura, Java se encarga de liberar la memoria del sistema.
- Políticas de seguridad.

- Independencia en la plataforma sobre la que se ejecute.
- El lenguaje permite el desarrollo de aplicaciones que puede ejecutar distintas líneas de código simultáneamente.
- Permite la compilación solo de la clase Java en la que se está trabajando. Java se encargará de buscar las clases asociadas para realizar la compilación completa.

5.4.3 JavaScript

El lenguaje de programación JavaScript [B.54] fue creado por Netscape, por la necesidad de facilitar el desarrollo de dinamismo en las páginas Web, sin la necesidad de utilizar scripts de CGI (del inglés *Common Gateway Interface*). Una de sus principales cualidades, es la multiplataforma, es decir, la posibilidad de implementación en todos los tipos de navegadores Web.

Este lenguaje trabaja como complemento del código HTML, permitiendo el desarrollo de aplicaciones a través de la Web. Este código, denominado *script*, se incluye dentro del HTML o bien en unos ficheros que posteriormente el HTML tendrá que incluir mediante una referencia.

5.4.4 JSP

Las páginas de servidor Java [B.55] (JSP, del inglés *JavaServer Page*) es una tecnología Java que nos ofrece la capacidad de generar contenido Web dinámico.

Las JSP nacen como una derivación de los *servlets* de Java, llegándose a considerar una alternativa a la creación de *servlets*. Cuando se ejecuta el JSP, el código es compilado y generad como si de una clase Java se tratase, pudiéndose ejecutar en la parte del servidor.

Sin embargo, la diferencia radica en la parte en la parte de programación. Un JSP contiene código etiquetado (por ejemplo, HTML) y código Java, mientras que el *servlet* es una aplicación por la que se recibe peticiones de consulta y se genera una respuesta en formato para páginas Web.

5.5 Base de datos

5.5.1 Introducción

Como se ha visto hasta ahora, todo sistema de recuperación de información necesita de una base de datos y programa de indización para gestionar y almacenar toda la información que vaya recuperando. Ante esto, durante el proyecto se barajaron varios sistemas de gestión de datos, en primer lugar se pensó en la utilización de sistemas basados en textos que devolviesen los documentos más relevantes para el usuario, como Lucene, sin embargo, tras el estudio de estas tecnologías y la presentación al cliente, se decantó por usar bases de datos racional, en concreto MySQL, el cual facilita la integración y simplifica el trabajo realizado en el proyecto.

Una base de datos se puede definir como una colección de información relacionada, que puede ser creada, actualizada y eliminada mediante un gestor de bases de datos (SGBD, del inglés *Database Management System*). Este tipo de gestores son muy específicos y están dedicados principalmente a ofrecer una interfaz entre el usuario y la base de datos.

Por lo tanto, una base de datos relacional puede ser definida como una base de datos atendida por tablas, columnas y filas que se relacionan entre ellas, basándose en unos valores claves contenidos en las

columnas.

En cuanto a la comunicación de bases de datos, encontramos el lenguaje de consulta estructurado (SQL, del inglés *Structured Query Language*). Este, fue creado por IBM como la interfaz del sistema de bases de datos relacional *System R*. Utilizando el SQL, se pueden crear órdenes para manejar los de los datos almacenados.

Actualmente, en el mercado existen muchos sistemas de administración de bases de datos, como MySQL, PostrgresSQL, Oracle...y además también existen estándares y frameworks que facilitan el acceso y la ejecución de operaciones sobre las bases de datos, como la API JDBC, ODBC o Hibernate.

5.5.2 MySQL

El sistema de administración de bases de datos relacional MySQL [B.56], es un sistema de código abierto y multiplataforma, que bien se puede utilizar mediante el interprete de ordenes (consola) o bien mediante una interfaz visual como *MySQL Workbench*.

Uno de los puntos más importantes de este sistema, y curiosamente no es tecnológico, es la gran cantidad de servicios en línea que ofrece para la ayuda al usuario. Todas las configuraciones, los comandos utilizados, curiosidades del sistema... pueden ser encontradas en su página Web, donde a través del foro o la ayuda presentada, el usuario puede consultar información y así resolver las dudas que tenga. Por otra parte, se caracteriza por ser un servidor que soporta información de forma masiva y que puede procesar consultas bastante complejas en un tiempo adecuado.

Otra característica, es la de *multitarea*, es decir, por cada usuario que establezca una conexión con el servidor, el gestor del servidor creará un subproceso para controlar la solicitud que ha recibido.

Por último, otro punto, y no menos importante, es la posibilidad de exportar las bases de datos a casi cualquier plataforma sin tener ningún tipo de problema.

5.5.3 Hibernate

Hibernate [B.57], es una capa de persistencia objeto-relacional (ORM, del inglés *Object-Relational Mapping*) para el lenguaje de programación Java. Este tipo de herramientas ayudan en el mapeo de las bases de datos relacionales, solucionando así el problema de la diferencia entre los modelos de datos que existen en una aplicación: la orientación a objetos y el modelo relacional de las bases de datos.

Para ello, el desarrollador puede puntualizar con cierta libertad todo el modelo de la base de datos. Mediante el uso de ficheros XML donde se generan los modelos de datos, se puede especificar cuales son las relaciones y que características tienen. Con esta información, Hibernate es capaz de operar en la base de datos, convirtiendo los datos utilizados por Java y a los definidos por SQL, liberando así al desarrollador de tener que usar sentencias SQL a la base de datos.

5.5.4 Características de MySQL como gestor de base de datos

Veamos a continuación algunos aspectos por los que se ha escogido MySQL.

- Es un sistema de bases de datos de elevado rendimiento, cuya configuración se hace menos compleja que otros sistemas más grandes.

- Es completamente portable a otras plataformas. De una forma sencilla y segura se puede exportar de una máquina a otra.

- Su seguridad y conectividad permiten la accesibilidad desde cualquier lugar de Internet.

- La conexión al servidor puede darse de forma simultánea por muchos usuarios.

- Su velocidad es una de sus principales características.

- Debido al uso de Hibernate en la aplicación, la compatibilidad de este resulta mejor con MySQL que con otros sistemas.

5.6 Almacenamiento masivo

Debido a que la aplicación tiene que estar preparada para soportar grandes cantidades de información, además de MySQL se han utilizado otro tipo de tecnologías para el almacenamiento masivo.

- Amazon S3: Es un servicio de Amazon que ofrece alojamiento de datos en la nube. Se caracteriza por la velocidad de transferencia de datos, por lo que la aplicación Cosmos la utiliza como almacenamiento de páginas Web ó imágenes.

- Lucene: Es un sistema de recuperación de información de código abierto desarrollado por Apache. Dispone de una versión en Java la cual es integrada a la herramienta Cosmos para realizar la indexación y búsqueda de información una vez recuperada de la red.

- Apache Solr: Es un motor de búsqueda de información que utiliza las librerías de Lucene. También es de código abierto, y es utilizado por la herramienta Cosmos para la integración de Lucene con el resto de tecnología.

6. MÓDULO DE RECUPERACIÓN

6.1 Estudio general de medios

Según el último estudio general de medios (EGM, Octubre 2011-Mayo 2012) [B.58], el cual uno de sus objetivos es la medición de la evolución y penetración de los medios de comunicación actuales, revela que Internet está en el quinto puesto (43.4%) en cuanto a la cantidad de individuos que lo utilizan.

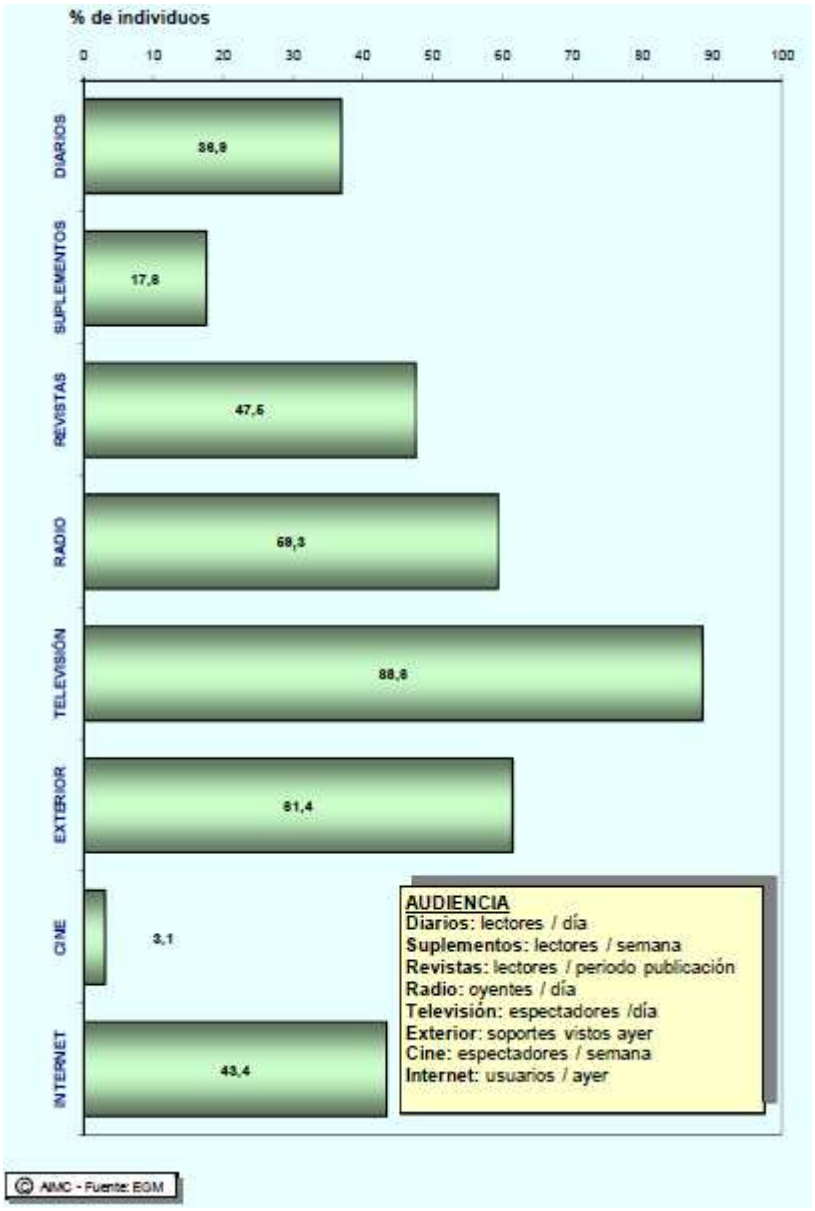

Figura 19. EGM

No obstante, si vemos la evolución de los medios de comunicación en los últimos años, a diferencia de algunos, Internet tiene un crecimiento exponencial conforme va pasando el tiempo, superando el nivel de penetración de otros medios.

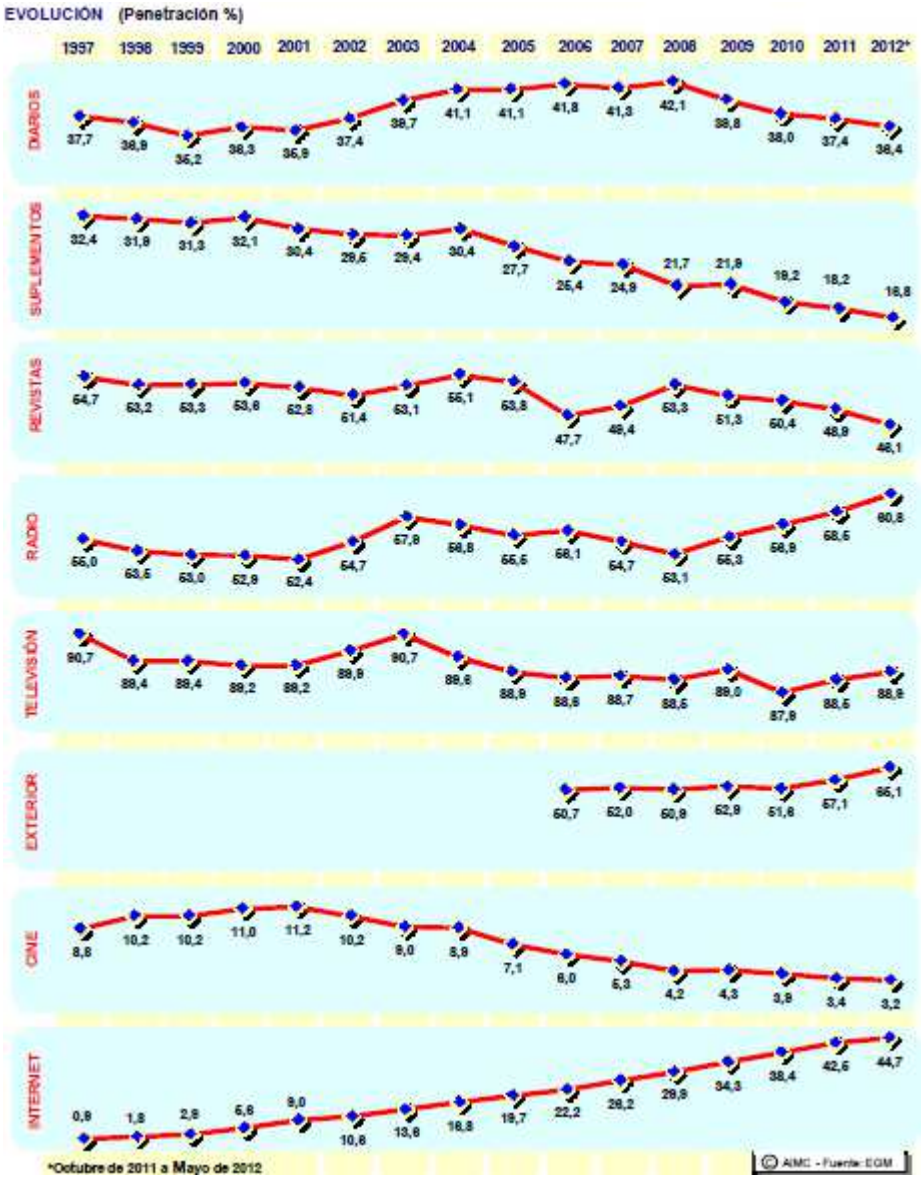

Figura 20. EGM: Evolución

Debido a este crecimiento y a la era digital en la que nos encontramos, los medios tradicionales se han visto obligados a redefinirse y adaptarse a este medio. No obstante, actualmente no son los únicos canales de información en Internet. Gracias a las evoluciones sufridas en la Web y al surgimiento de nuevas herramientas colaborativas, han provocado la aparición de nuevos canales en la red.

Actualmente en Internet se pueden encontrar infinidad de canales de información que son de gran interés para la escucha activa, desde Webs estáticas que aportan información en un momento dado hasta las redes sociales donde la información fluye constantemente.

Debido a su naturaleza este presenta algunas peculiaridades propias del medio y que a su vez son comunes para todos los canales de información a los que engloba.

La cualidad hipertextual de las páginas Web, ofrece al usuario realizar una lectura no secuencial del contenido y permitiéndole la libertad de desplazarse por la página e ir buscando lo que realmente le resulte interesante para él.

La globalización ha permitido que la información pueda ser visualizada prácticamente a nivel mundial. Ésta ha pasado de ser un recurso en algunos casos privados a uno completamente público y abierto. Si además, añadimos que la información en Internet en muchos casos no es dependiente del tiempo, pudiendo encontrar contenidos nuevos o antiguos, descubrimos una propiedad de asincronismo en este medio. Así pues, debido a la posibilidad de almacenar la información en Internet, muchas veces este medio se convierte en una gran base de datos en la que la información es conservada en el transcurso del tiempo.

Con la evolución de la Web y sus diferentes formatos de programación han permitido la integración de elementos multimedia en las páginas. Así pues, Internet reúne muchas de las características de multimedialidad de los medios de comunicación tradicionales y aporta muchas más posibilidades de explotación de estos elementos.

Actualmente, una de las características de las Webs 2.0 es la interactividad, por lo que los canales en este medio permiten que los usuarios participen con la página, dejando de ser meramente consumidores de información y convirtiéndose en receptores participativos de la comunicación.

Si dejamos a un lado la mera digitalización de los medios de comunicación tradicionales, vemos que la característica más importante reside en la participación del usuario con los nuevos canales, permitiéndole profundizar en los contenidos y escoger con total libertad a la información de interés.

Cabe diferenciar los medios de comunicación masivos tradicionales de lo que es Internet actualmente.

En los medios tradicionales el emisor se dirige habitualmente a muchas personas, que en su mayor medida no se conocen y se engloban en lo que se considera el público. Sin embargo, en Internet el emisor normalmente se dirige a una cantidad de usuarios inidentificables, ya sea por sus perfiles o por otras propiedades.

Los medios tradicionales suelen estar limitados a una zona geográfica y en un acotado período de tiempo, mientras que debido a la globalización de internet y a su característica de memoria, permite que la información sea accesible desde cualquier parte del mundo, sin limitaciones de tiempo.

Además, aunque existen avances por los cuales se permite interactuar las personas con los medios tradicionales, las posibilidades que ofrece Internet dentro de este ámbito superan con creces al de los cotidianos.

6.2 Fuentes de información del proyecto

Como se mencionó anteriormente, en Internet se pueden encontrar infinidad de canales de información, y es por ello, que se hace necesario la clasificación de estos canales según la naturaleza de donde provenga la información.

A grandes rasgos, en la escucha activa de Cosmos podemos encontrar los canales textuales, que engloban medios como la prensa, los *blogs*, los foros y en general la Web, los canales multimedia donde podemos encontrar plataformas como YouTube o Flickr, los canales de microblogging como Twitter o los canales de redes sociales como Facebook.

Gran parte de las páginas Web se consideran estáticas o con una frecuencia de actualización muy baja. Entre ellas, se destaca las páginas estilo enciclopedias, las páginas corporativas de algunas empresas, algunas páginas personales…en definitiva, engloba páginas que son revisadas con poca periodicidad a lo largo del tiempo. Sin embargo, existen otros tipos de Webs que se suelen actualizar continuamente o con una brevedad no muy alejada a la de un día. Dentro de lo que consideramos el canal textual, además de la propia Web como tal, este engloba distintos ámbitos según la utilización: la prensa online, los *blogs* y los foros.

A diferencia de los medios tradicionales como la prensa y en la actualidad en sus evoluciones online, existe una comunicación alternativa, una con una filosofía diferente, una enlazada con el desarrollo de la sociedad, una personal y a la vez participativa…con el único fin de fortalecer y dar a conocer conocimiento libre en este mundo globalizado.

Esta variante de comunicación se puede observar en los *blogs* y en los foros, donde el usuario expresa su visión de lo que le rodea resultando un contenido que en muchas ocasiones se ve contradicho con los medios tradicionales supervisado por el discurso de algunos poderes.

Otro fenómeno democratizado y que está teniendo gran impacto en la sociedad es el *microblogging* o *nanoblogging.* Éste permite a los usuarios mandar y recibir mensajes de texto cortos. Esta peculiaridad propia del canal ha favorecido a que se convierta en una de las más importantes fuentes de generación de información, llegándose a considerar como el canal de información de última hora. Su naturaleza de contenido reducido y su sencillez de uso han sido cruciales para el éxito de este tipo de canales. El impacto que ha suscitado en la sociedad ha provocado que se vea incrementado el número de usuarios activos de forma exponencial. La comunicación bidireccional que ofrece y la circulación constante de información de actualidad son algunos de los puntos clave por los que este canal se considera como una fuente de información en tiempo real. Y es por ello, que al ser un canal de actualidad donde los usuarios cuentan los últimos sucesos que están ocurriendo, se convierte en una de las piezas principales de la escucha activa.

Con la llegada de las nuevas tecnologías y la necesidad de las personas por crear sistemas de grupos o comunidades, surgen las redes donde los usuarios mantienen diferentes relaciones entre ellos con una afinidad en común. Las redes sociales han pasado a ser una costumbre para la nueva sociedad. El uso de estos canales de información ha transformado los estilos de vida de las personas así como las estrategias seguidas por algunas organizaciones y empresas. Las posibilidades que residen en estas aplicaciones, la comunicación inmediata y la cantidad de información que se genera a diario, hacen que este medio sea de diversa utilidad en distintos ámbitos, entre los que destacamos la escucha activa.

La aparición de nuevos mecanismos digitales multimedia como las cámaras fotográficas, las de vídeo, los teléfonos móviles…y el abaratamiento de estos dispositivos, está favoreciendo a que las personas almacenen grandes cantidades de información digital. Si a esto, le sumamos la aparición de los servicios de intercambio de contenidos multimedia en Internet, surgen nuevos canales, los cuales se caracterizan por almacenar una gigantesca base de datos de contenidos multimedia aportados por los usuarios y permitiéndoles su difusión e intercambio de los mismos. Así pues, de este tipo de canales destacamos dos: YouTube y Flickr. En el primero, los usuarios pueden compartir vídeos, mientras que en el segundo, pueden compartir tanto vídeos como imágenes.

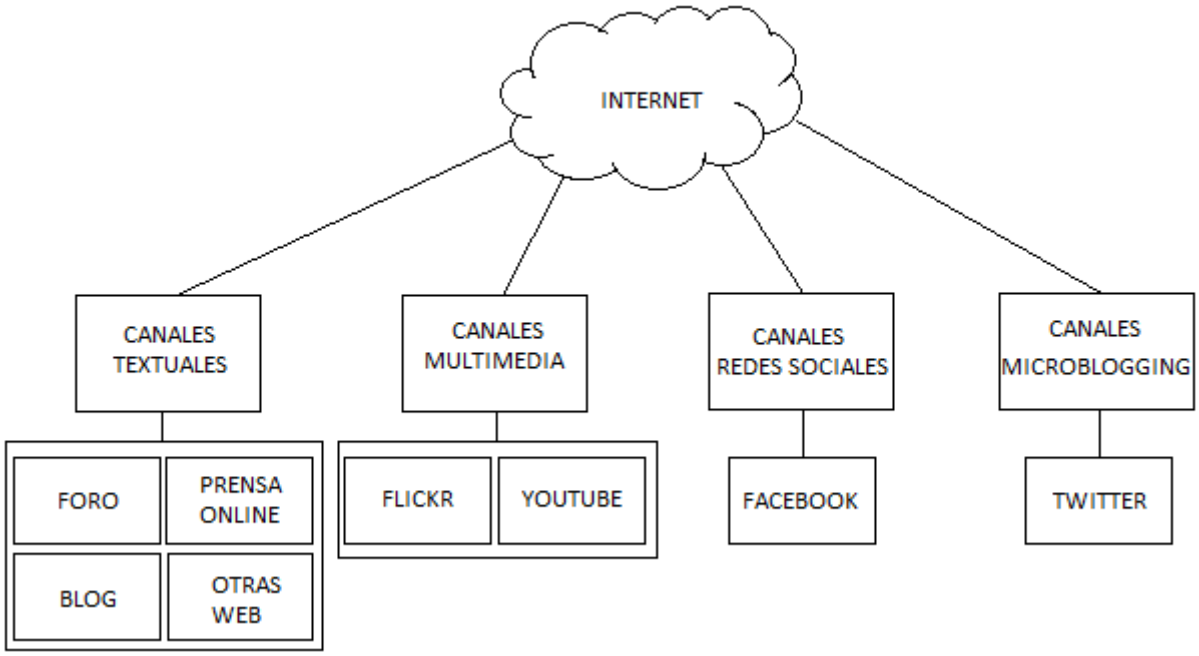

Figura 21. Fuentes de información

6.3 Retos y estrategias de la recuperación de información

Ante la naturaleza de Internet y las tendencias asociadas a los sistemas de recuperación de información encontramos algunos retos a los que ha de hacer frente el sistema. A continuación veremos los retos establecidos para el sistema de recuperación de información y como han sido abordados.

El primer reto hace referencia a la eficiencia. En el mayor grado posible, el sistema ha de ser lo más rápido posible en la recuperación de información y obtener la máxima cantidad de resultados.

El segundo reto es la calidad de la información obtenida. La relevancia de la información es un punto crucial para el usuario que realiza la consulta, por lo que el sistema ha de ir aprendiendo e ir descartando la información que el usuario ha indicado como no relevante.

El tercer reto es la paralelización. Debido a que la herramienta va ha soportar grandes cantidades de proyectos se tiene que abordar una estrategia de paralelización de las máximas tareas posibles.

El cuarto reto es el almacenamiento de la información. El sistema ha de recuperar la información y después almacenarla en bases de conocimiento. Ha de tener una buena gestión debido a que las cantidades de información posiblemente sean desorbitadas.

El quinto reto hace referencia a la usabilidad. El sistema ha de presentarse lo más amigable para el usuario, facilitándole el proceso de recuperación durante todo momento.

En cuanto a las estrategias seguidas en el proyecto de recuperación de información se pueden dividir en tres modelos.

La primera se centra en la utilización de APIs ofrecidas por las herramientas de la Web 2.0. Con esto, nos referimos a las aplicaciones que a través de ellas se pueden utilizar los recursos que ofrecen las herramientas para obtener la información. No obstante, como se verá más adelante este tipo de interfaces van

ligadas a una serie de restricciones impuestas por el propietario de la herramienta, y que presentan una serie de desafíos en su implementación.

La segunda se basa en la utilización de *crawler* clásicos y específicos. Como se verá a continuación este tipo de herramientas presenta algunas ventajas e inconvenientes frente a las interfaces de programación, de las que destacamos la captación de mayor cantidad de información.

La tercera, y última estrategia seguida es la combinación de ambas. Mediante la utilización de las dos estrategias anteriores se puede mejorar la información obtenida por cada una de ellas por separado, de esta forma se precisa de mayor cantidad de información para que el usuario tenga un pleno conocimiento de los contenidos recuperados.

6.4 Fuentes textuales: Nuevos medios

6.4.1 Canales de información

De la necesidad de libertad de expresión por parte de los usuarios y de las continuas censuras que se pueden apreciar en la prensa digital nacen los *blogs*. Un blog o también llamado *weblog* es considerado como una bitácora personal, es decir, un sitio donde un usuario escribe contenidos de temas que le resulten interesantes, siempre ordenados de forma cronológica. Los *blogs* nacen de la necesidad de expresión por parte de los usuarios, de la libertad de comunicar sus experiencias e inquietudes, del periodismo ciudadano…Al contrario que ocurre en los grandes medios de prensa digital, los *blogs* expresan el punto de vista de lo personal, de la participación individualizada y quizá desde una perspectiva más objetiva. Es la naturaleza del individuo, que a través de los contenidos expresa sus inquietudes, ideas y experiencias sobre un tema en concreto, resultando en un lugar de referencia para muchos usuarios y formando así una comunidad que comparte un propósito en común.

A diferencia de otros canales, los *blogs* no tienen una definición contextual inicial, sino que es el propietario de la bitácora quien definirá cual será su uso y para que fines, estableciendo así un tema general para todo el sitio que perdurará en el tiempo. Esta es una de las propiedades que definen a un *blog*, ya que no consiste en exponer información abarcando infinidad de temas, sino que el contenido está centrado en uno solo.

La confianza subyacente de los *blogs* viene remarcada por una serie de aspectos éticos que han ido surgiendo de la familiaridad con la que se comunican y desarrollan comunidades de usuarios en torno al sitio. La aproximación de la comunicación con los usuarios, el estilo de escritura más directo, las conversaciones más cercanas, el reconocimiento al equivocarse y el vínculo a otras páginas (*blogroll*) o citas, son algunos de los aspectos de porque este tipo de canal tiene un gran impacto en la sociedad, viéndose incrementado el número de páginas en las que se produce un periodismo ciudadano alejado de los formalismos de la prensa digital.

Así pues, los *blogs* se consideran fuentes de información democráticas de gran importancia en el ámbito periodístico y de la información. Las grandes cantidades de *blogs* que podemos encontrar en la Web, y algunos de muy buena calidad, ayudan en el intercambio de conocimiento entre los usuarios, siendo uno de los puntos de referencia de los principales periódicos digitales y de comunidades de usuarios con un tema en común. La gran importancia que tienen como contribuidores de información representa una de las fuentes que no puede ser ignorada.

Otra de las fuentes textuales son los foros. La idea de estos trasciende de los comunes foros romanos. En estos sitios de interés, se reunían las personas para debatir ciertos temas. Así pues, esta filosofía se trasladó a la era digital, dando forma a lo que conocemos actualmente como foros online, la forma más madura de

herramienta social con el objetivo de fomentar el dialogo entre muchos usuarios en una página Web.

Los foros son los descendientes de los sistemas de noticias *BBS* (del inglés *Bulletin Board System*), y normalmente van ligadas a una página Web. En ellos se producen grandes cantidades de discusiones y opiniones alrededor de un tema o subtema de interés, incitando al usuario a que participe en una discusión libre e informal.

En estos canales existe claramente una jerarquía social, en la que se pueden encontrar diferentes roles según la función de cada usuario:

- Los administradores, son los encargados de gestionar y supervisar el portal. En ellos reside toda la responsabilidad del sitio. Se encargan de marcar una serie de normas para el sitio, los temas a tratar y podrán designar diferentes roles a usuarios estableciéndoles una serie de funciones a cometer.

- Los moderadores, son los usuarios que a través de adjudicación de funciones por el administrador, se encargan por lo general, de la modificación o eliminación de contenido, del traspaso de unos contenidos a otros…tareas de mantenimiento con el único objetivo de mantener un aspecto más amistoso y organizado.

- Los usuarios, únicamente tienen la posibilidad de consumir o aportar información entorno a un tema.

La estructura de un foro viene determinada por los temas que se van a tratar. El administrador se encarga de establecer una serie de temáticas en los que da la libertad de expresión (siempre que se cumplan las normas) a los usuarios. A su vez, el administrador puede establecer una serie de subtemas con el objetivo de organizar y controlar mejor el portal.

Dentro de los tipos de foros se pueden encontrar tres:

- Públicos: son aquellos en los que los usuarios pueden participar sin tener que registrarse.

- Protegidos: son aquellos que para dejar un comentario es necesario registrarse.

- Privados: son aquellos en los que solo se puede interactuar con ellos si el administrador del foro acepta la solicitud de ingreso.

Debido a la versatilidad que presentan, los foros se han convertido en un espacio de reunión donde los usuarios pueden debatir en torno a un tema de interés, facilitando el contacto entre personas con aficiones comunes. A pesar de que la comunicación dentro de los foros no es a tiempo real, el flujo de información suele ser fluida. Así pues, sirven como comunidades de apoyo a páginas Web para facilitar la continuidad y participación de los usuarios en esta.

Una de las diferencias que cualifica a este canal, es que el foco de información no está dirigido a un tema en particular ni concreto, sino a una visión más general y que abarca todos los temas que haya considerado oportuno el administrador del sitio.

Otra característica de los foros, es la posibilidad de categorizar los temas a debatir en una serie de grupos o contenedores. Además, en estas herramientas se permiten un elevado número de usuarios y los temas son mucho más diversos y extensos. La discusión sobre un tema se encuentra de forma anidada y ordenada cronológicamente, asemejándose a los comentarios de los *blogs*.

Normalmente, este tipo de canal se comporta como un sistema de pregunta y respuesta. Los usuarios aportan contenidos, que en la mayoría de ocasiones son peticiones, problemas o dudas que necesitan resolver. Así pues, a través de la creación de un debate por parte del usuario, la comunidad puede ir aportando sus opiniones y respuestas en señal de ayuda.

Sin embargo, con la aparición de las redes sociales y otros canales de información, los foros se están viendo eclipsados por estas nuevas tendencias. La capacidad de anonimato, el mal uso que algunos les dan, la influencia de las redes sociales así como la cantidad de usuarios que debaten entorno a este canal, la dilatación del tiempo en las conversaciones…son algunos de los aspectos de porqué los foros están tendiendo a migrar a las redes sociales y dejar un poco apartado los sistemas tradicionales.

No obstante, la integración de foros en las páginas Web sigue siendo algo cotidiano y habitual, aunque la tendencia está provocando que estos se estén dirigiendo a las redes sociales. Esta evolución de los foros, está permitiendo que en las redes sociales haya zonas especializadas o grupos de interés en los que los usuarios con un tema en común puedan aportar contenidos de una forma más interactiva.

Así pues, un foro es considerado como un canal donde se debaten temas o subtemas relacionados, a través de la interacción de diversos usuarios, en un tiempo diferido.

Por último, dentro de estas fuentes de información encontramos lo que consideramos como otras Webs, es decir, si descartamos lo que son el resto de canales de información (redes sociales, foros, *blogs*, diarios…) y nos quedamos con lo que queda, surgen páginas Web, que por lo general tienen poca actividad y su periodicidad de actualización es reducida. Dentro de este conjunto de páginas Webs, encontramos algunas como las personales, sitios webs para profesionales…

La Web es la forma más básica de representación de un documento en Internet. Está compuesta generalmente por información en texto, hiperenlaces y elementos multimedia. Además puede contener estructuras de datos de estilos dándole un aspecto visual más atractivo. Normalmente son escritas en formato HTML, aunque también existen otros formatos con los que implementarlas. Las páginas Web se definen en: estáticas y dinámicas.

Las páginas Web estáticas son consideradas todas aquellas Webs que no son actualizadas frecuentemente, además de no ofrecer la posibilidad de interactividad por parte del usuario. Al contrario ocurre con las páginas Web dinámicas, donde el usuario puede participar de forma activa con los contenidos, además, el proceso de actualización de estos se hace de forma mucho más sencilla, por lo que favorece a que la periodicidad de agregación de nueva información vaya cambiando frecuentemente.

Al igual que en muchas de las fuentes de información, en una página Web se pueden incluir tanto texto como elementos multimedia, en cambio, no está pensado para la interacción con los usuarios, sino que se mantienen unos contenidos estáticos con la finalidad única de informar a los usuarios. Así pues, se rompe totalmente con el vínculo hacía el usuario.

6.4.2 Ámbitos de clasificación

Centrándonos en la identificación de los *blogs*, según el estudio realizado por Francisco Manuel Rangel y Anselmo Peñas [B.59] la identificación de este tipo de fuentes se puede obtener a partir de una serie de características visuales que los hacen identificables por un humano a simple vista, estas son:

- Bloques de información, denominados posts, que están estructurados como entradas de un diario, con una fecha, un título, un contenido, y la posibilidad de introducir comentarios por los lectores
- Un *blogroll* o grupo de links proporcionando enlace permanente a contenidos clasificados del propio blog

- Palabras altamente representativas por su alta frecuencia de aparición como son blog, post, RSS, Atom, comentario...

Estas características visuales se modelan mediante una serie de atributos identificables componiendo una representación basada en los *frames* de Minsky que sirven de base para el entrenamiento y posterior validación con un clasificador binario.

Si se aprecian los resultados obtenidos, se puede afirmar que con el análisis de la página centrado en estas tres características visuales se puede discriminar entre lo que es un *blog* y lo que no es, casi con total certeza.

Pero a diferencia de los sistemas de clasificación automática que en muchos aspectos sirven como sugerencia para un equipo de analistas, en este tipo de proyectos en el que el usuario final es mucho más critico y los resultados dependen inevitablemente del éxito o el fracaso del producto, se ha de aproximar los resultados obtenidos a una tasa de acierto perfecta. Por ello, se utiliza el clasificador en una fase de *backoffice* que sugiere a los analistas qué documentos son *blogs* y qué documentos no lo son, asistiendo en la creación de una enorme base de datos de inteligencia (BDIA: Base de Datos de Inteligencia de Autoritas) que servirá como base para la clasificación de este tipo de páginas.

Para clasificar foros se podría imitar la estrategia de la clasificación de *blogs*, pero por limitación de recursos y porque la cantidad de foros relevantes es inferior, se ha optado por una clasificación manual en base a análisis de directorios especializados por un especialista.

En caso de la clasificación en un foro o en un blog, se determina que la página es de un tipo considerado como otras Webs.

6.4.3 Particularidades de la recuperación

Cada una de estas fuentes presenta algunas características en común frente a la recuperación de información, sin embargo otras son específicas dependiendo de la fuente.

La periodicidad de agregación de nuevos contenidos, así como la interacción por parte de los usuarios son características que definen o deberían definir a toda página Web. La importancia de actualización y agregación de nuevos contenidos es uno de los pilares fundamentales para el éxito de una página y que pueda perdurar durante el tiempo. No obstante, al ser responsabilidad del usuario el mantenimiento y actualización del sitio, muchas veces los contenidos trascienden en el tiempo sin verse modificados, quedando en el olvido para muchos usuarios de la comunidad.

Un factor primordial que afecta en la recuperación, es la cantidad de información devuelta por las APIs. Al hacer uso de estas, la descripción del contenido viene acotada por unas líneas compuestas por aproximadamente 300 caracteres, siendo en ocasiones insuficiente información para comprender el contenido. Así pues, además de las APIs de búsqueda se precisa de un *crawler* específico que obtenga toda la información del contenido de la página.

Para conocer mejor y verificar la información de la fuente, es importante identificar al autor y así relacionarlo con el contenido, averiguar sus destrezas, conocer sus críticas…en resumen, conocerlo a él y así poderlo identificar. Sin embargo, esto no siempre es posible en algunas de las fuentes citadas. Una característica que se da en muchos foros, es la no identificación de los usuarios. En muchos foros se da la posibilidad de escribir de forma anónima por lo que se hace imposible obtener el perfil del usuario. Además, aunque se precise de un registro previo para participar, se considera un canal menos identificativo y más oscuro, donde muchos de los usuarios quieren permanecer en el anonimato, por lo que introducen en su gran

mayoría identificaciones falsas que poco se asemejan a la realidad. Al contrario pasa en los otros dos canales de información, donde el propietario puede ser identificado dentro de los contenidos de la página.

Otra peculiaridad que afecta a la recuperación de información, es la fecha de los contenidos. A diferencia de otras implementaciones, la fecha obtenida por las APIs de recuperación de información indica la última vez que el *crawler* visitó la página, por lo que en muchas ocasiones no se corresponde con el día en que se publicó el contenido. Este problema persiste para todos los canales textuales menos en las consideradas como otras Webs, y es que, en esta ocasión, el *crawler* de búsqueda favorece al devolver la fecha en la que se realizó la última actualización de la página. Una posible solución para solventar este problema, seria hacer un *crawler* tradicional que consultase la página y obtuviese todo el contenido en su formato original, siempre y cuando este permitido esta recuperación por parte del sitio Web.

Como se ha visto anteriormente, otra característica de estas fuentes es la privacidad que puede ofrecer, haciéndose imposible la recuperación si el contenido de la página es privado.

Otro problema que podemos encontrar en este tipo de páginas es que sufren de varios enemigos claramente definidos:

- El *spam*. Existen grandes cantidades de información que son publicaciones de mensajes no aceptados ni solicitados, y que en gran medida únicamente aportan contenido publicitario dificultando el correcto funcionamiento del canal. Esta tarea es responsabilidad del administrador de ir controlando este fenómeno para que en la medida de lo posible pueda atajarse sin generar nuevos problemas.

- Los *troll*. Son usuarios que únicamente tienen como finalidad romper con el ambiente cordial y familiar que pueda haber en un foro, generando nuevos contenidos donde expresa su desacuerdo con las temáticas seguidas o simplemente por molestar.

- Los *chaters*. Son usuarios que no escriben gramaticalmente correcto, y que sus contenidos se asemejan mucho a las escrituras utilizadas en los SMS de móviles, por lo que dificultan en gran medida la lectura y comprensión de la información.

Por último, otra característica que podemos encontrar es la diferenciación de los contenidos y los comentarios. Debido a que las APIs únicamente devuelven un pequeño fragmento del contenido y una parte interesante son los comentarios de los usuarios sobre la información plasmada, se hace necesario analizar los contenidos obtenidos en su formato original para buscar y discriminar lo que es un comentario de lo que es el contenido de la información.

6.4.4 Recuperación: Información y limitaciones

Para recuperar información de este tipo de fuentes podemos encontrar tres alternativas: uso de las APIs de los buscadores, un *crawler* clásico o un *crawler* guiado por RSS, que a su vez pueden ser complementarias. Como el proyecto se centra en la utilización de APIs para la recuperación de información, se ha decantado por la implementación de las APIs de los buscadores más comunes para obtener a partir de ellos la información de estas fuentes, no obstante debido a las problemáticas de la poca descripción que devuelven las APIs, también se ha implementado un *crawler* clásico, que a partir de las URL obtenidas por los resultados devueltos por las APIs, el *crawler* obtiene la información de la página.

6.4.4.1 Google

Para utilizar la interfaz [E.28] se ha implementado un servicio que realiza peticiones *RESTful* (del inglés *Representational State Transfer*), al método que proporciona la interfaz de aplicación de Google. Con la utilización de esta interfaz, el sistema puede realizar las peticiones de consulta por parte del usuario igual como si se tratase de la página Web del buscador. El servicio ofrece un método, el *customsearch*, el cual devuelve los resultados de la consulta solicitada por el usuario.

No obstante, la API de Google necesita de una clave (*APIKey*) para poder hacer uso de la interfaz. Para obtenerla, es necesario contar con una cuenta de correo de *Gmail* y registrar una aplicación de desarrollador, que al darla de alta nos proporciona la clave de acceso para realizar las peticiones de consulta.

En cuanto a las restricciones, la API proporciona 100 consultas diarias de forma totalmente gratuita, teniendo la posibilidad de aumentar hasta 10000 peticiones por un coste de 5 dólares al mes.

Los resultados pueden ser mostrados tanto en JSON como en Atom. A continuación se muestra una tabla con los parámetros utilizados para realizar las consultas a la interfaz:

Resultado	**Descripción**
Key	API Key requerido para realizar las peticiones.
Fields	Selección de un subconjunto de campos incluidos en las respuestas.
Lr	Idioma en que queremos obtener los resultados. Los disponibles son: Inglés, Francés, Italiano, Español, Alemán y Portugués.
Q	Indica la consulta por parte de los usuarios.
Start	Indica el primer resultado a devolver.
Num	Indica el número de resultados a devolver.
Gl	Indica de qué país han de ser los resultados obtenidos. El valor de parámetro viene indicado por dos letras del código del país.

Figura 22. Google: Tabla de resultado

Debido a que la interfaz de Google no proporciona la fecha del documento, para una primera aproximación se utiliza el día en que el documento fue recuperado por nuestro crawler.

6.4.4.2 Yahoo!

Para la implementación del consumo de la interfaz de Yahoo! [E.29] se ha realizado un servicio que consume las peticiones a la API a partir de llamadas *RESTful* al método de búsqueda proporcionado por el servicio.

Al igual que la otra interfaz, Yahoo! precisa de la obtención de una APIKey que verifique la identidad del sistema. Así pues, con una cuenta de correo de Yahoo!, y el registro de una aplicación de desarrollador en la página Web de la API, se obtiene la clave de acceso. No obstante, para verificar la autentificación del usuario, es necesario realizar una autentificación OAuth por parte del sistema.

La API proporciona diferentes fuentes de información donde realizar la consulta por parte del usuario, destacando entre ellos la posibilidad de hacer la consulta en la Web, en imágenes, en noticias y en *blogs*, aunque este último se encuentra en versión Beta.

En cuanto a las limitaciones, el servicio restringe la cantidad de peticiones a 5000 consultas por mes en la aplicación registrada.

Los resultados pueden estar expresados tanto en XML como en JSON. A continuación se muestra una tabla con los parámetros utilizados para realizar las consultas a la interfaz:

Parámetros	**Descripción**
Service	Debe de especificar que tipo de servicio se va a utilizar: Web, Noticias, Imágenes o *Blogs*.
Market	Especifica el país de donde obtener los resultados.
Count	Indica el número de respuesta a mostrar por página. Como máximo es 50, excepto 20 para los *blogs* y 35 para las imágenes.
Language	Identifica el lenguaje a utilizar para obtener los resultados.
Format	Especifica el formato en que se presentan los resultados.
Sites	Se puede restringir el sistema a una determinada colección de Webs específicas.
Start	Indica el primer resultado a mostrar.
totalresults	Indica el número total de resultados. Como máximo son 1000.
Results	Por defecto son 10 resultados por página, pudiéndose escalar hasta los 1000 resultados posibles.
Adult	Filtro para la moderación de páginas con contenido para adultos.

Figura 23. Yahoo!: Tabla de resultado

6.5.4.3 Bing

Para el consumo de las peticiones proporcionadas por la API de Bing [E.30], se ha implementado un servicio que mediante peticiones GET a una URL donde se encuentra el servicio de la API, el método implementado obtiene los resultados de la búsqueda proporcionada a la interfaz, destacando esta por los resultados fuertemente tipados.

Al igual como en el resto de interfaces, esta API precisa de una verificación para su utilización. Sin embargo, aquí varía un poco al resto, ya que no es una APIKey, sino el ID de la aplicación de desarrollador que se tiene que registrar. Por lo tanto, se hace necesario tener una cuenta de correo de Microsoft y posteriormente ingresar en la página Web de la API y registrar una aplicación de consulta de información.

A parte del parámetro *AppID* es obligatorio dos parámetros más, la consulta y los *sources*. La consulta es el texto a buscar, mientras que los *sources*, es uno o más valores separados por el símbolo + y que indican el tipo de fuentes a utilizar en la petición de consulta. El API permite la búsqueda de información en muchas fuentes, de las cuales hemos escogido tres para nuestro proyecto: en la Web, en imágenes y en noticias.

En cuanto a las limitaciones, el servicio restringe la cantidad de peticiones a 5000 consultas por mes en la aplicación registrada. No obstante, esta cantidad puede ser incrementada si se paga una cuota mensual en función de las peticiones que se desee.

Los resultados obtenidos son presentados en formato JSON. A continuación se muestra una tabla con los parámetros utilizados para realizar las consultas a la interfaz:

Parámetros	**Descripción**
Adult	Filtro para la moderación de páginas con contenido para adultos.
Market	Especifica el lenguaje y la región de la petición.
Version	Indica a que versión de la API hace referencia: 2.0 o 2.1
Latitude	Para indicar la latitud de una petición.
Longitude	Para indicar la longitud de una petición.
Count	Indica el número de resultados por petición. Se puede indicar un valor entre 1 y 50.
Offset	Indica el primer resultado a mostrar. Se puede indicar un valor entre 1 y 1000.
Source	Específica de fuente queremos obtener la información.

Figura 24. Bing: Tabla de resultado

6.5.4.4 Crawler clásico

El crawler esta diseñado para que a partir de una URL dada, sea capaz de obtener la información HTML de una página de forma automática. Su principal objetivo es obtener la información que mediante las

APIs textuales no es posible, con ello, nos referimos principalmente a adquirir en la medida de lo posible todo el contenido de la página.

Así pues, en el momento en que la información es recuperada por las APIs de búsqueda, el sistema obtendrá la URL del sitio, y realizará una petición a la página para obtener su contenido HTML.

6.4.5 Implementación en Cosmos

Para la implementación del *crawler* de estas fuentes en la aplicación de escucha activa de Cosmos se ha construido un sistema compuesto por diferentes módulos.

El primer módulo es el encargado de obtener la información a partir de las APIs de los principales buscadores. Mediante la implementación del código que permite realizar las peticiones a estas interfaces, el sistema obtiene la información devuelta por los buscadores para su posterior análisis. Como se ha visto anteriormente, la información que define el contenido de la página devuelta por las APIs de los buscadores no es suficiente, por lo que se ha realizada un segundo módulo por el que se obtiene todo el contenido de la URL proporcionada por la API.

Dentro de este módulo y antes de limpiar todo el contenido del HTML, se realiza el clasificador de la página Web según el modelo aportado por los autores del estudio. De esta forma, la página es clasificada en un tipo de fuente propuesta en la recuperación de información.

Llegados a este punto, el sistema ya conoce la naturaleza de la fuente, por lo que el HTML es filtrado y tratado para exponer únicamente la información relevante para el usuario. Para ello, se utilizan técnicas de extracción de información que permiten discriminar la estructura y los contenidos no relevantes de los que sí lo son.

Por último, una vez analizado todo y extraída la información relevante, esta es almacenada en una representación textual comprensible para el usuario final.

El desarrollo de estas técnicas de recuperación de información y refinamiento permitirán tratar con la información masiva.

En la siguiente figura se muestra las partes funcionales que se siguen en la obtención de información de estas fuentes.

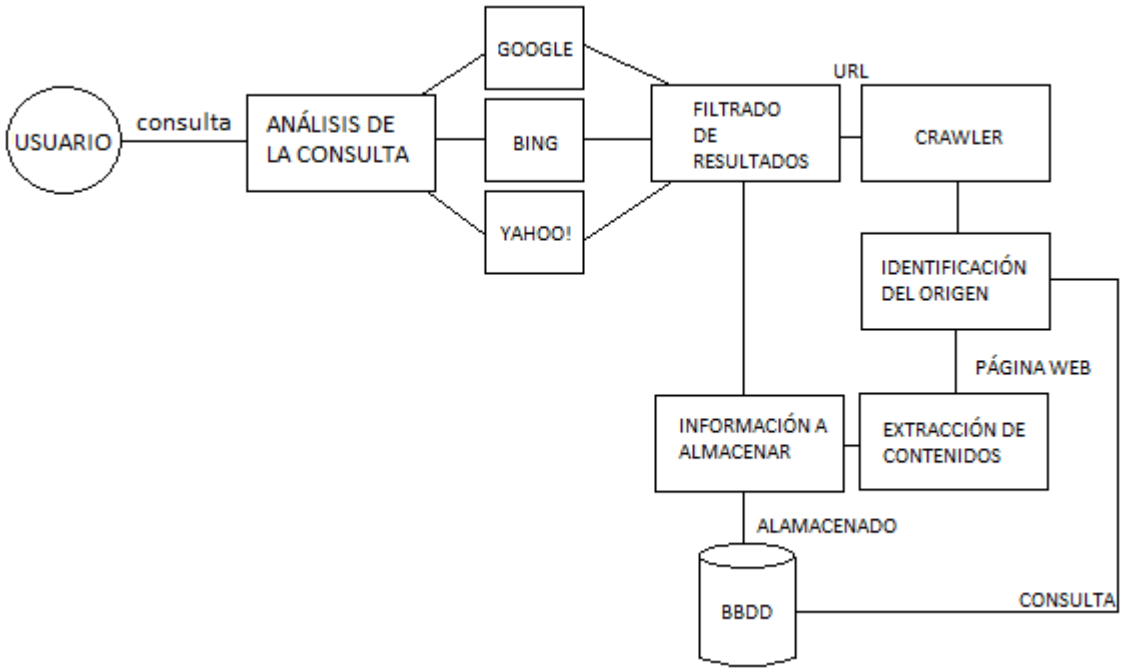

Figura 25. Proceso de recuperación en nuevos medios

Inicialmente el usuario realiza una consulta de información al sistema, el cual analizará dicha consulta y sus filtros correspondientes para ajustarla a las diferentes interfaces de programación. Una vez analizada la consulta, se distribuye a las diferentes APIs implementadas, las cuales devolverán los resultados correspondientes a la consulta del usuario.

Cuando se tienen los resultados, estos son filtrados para analizar su posterior clasificación según su naturaleza. Para ello, de los resultados obtenidos se pasa la URL al *crawler* que se encargará de obtener la información por completo de la página Web.

Una vez obtenido el código de la página, se consulta en la base de datos si el *host* de la página Web es conocido. En el caso de ser conocido, la página ya esta identificada. Por el contrario, se utiliza el algoritmo basado en características visto anteriormente para la categorización de si una página es un blog o no lo es.

En ambos casos, cuando la página es categorizada se extrae toda la información relevante, discriminando toda la parte de código y la información no relevantes, quedando al final, un texto plano que sea comprensible para el usuario.

Por último, la información obtenida de las APIs junto con la extracción de contenidos de la página Web es almacenada en la base de datos de conocimiento para su posterior análisis en los distintos módulos de la escucha activa.

Este modelo serviría para una ejecución secuencial, sin embargo, la realidad es que el sistema ha de soportar diferentes proyectos dando respuestas en un tiempo adecuado, así que para asegurar este procesamiento paralelo de altas prestaciones se ha optado por un modelo basado en *MapReduce* y mecanismos de sincronización por colas, definiendo la siguiente arquitectura:

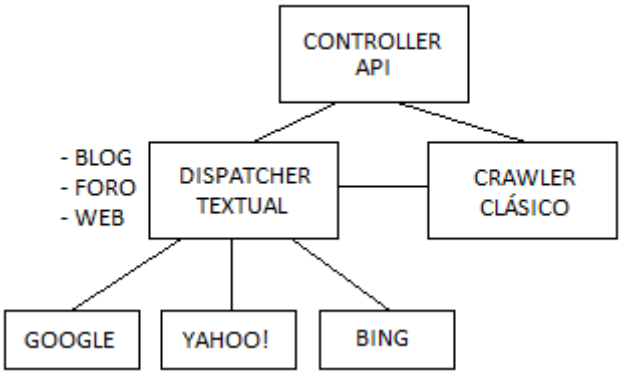

Figura 26. Arquitectura en nuevos medios

El *controller API* se encarga de gestionar todas las consultas que existan por parte de los usuarios. La consulta es enviada al *dispatcher textual* que las analizará y las transformará según a la API de recuperación de información que se corresponda. Los datos son recogidos por las APIs y devueltos de nuevo al *dispatcher*, donde serán tratados y filtrados. Como se ha visto anteriormente, para captar la información de los contenidos se hace necesario un módulo adicional por el cual se obtenga todo el contenido de la página Web, por lo que el *dispatcher* envía las URL de los resultados al *crawler* clásico que se encargará de recuperar y analizar todo el contenido de la página. Una vez filtrado el resultado, el *crawler* envía el contenido de la página en su forma normal, para que el *dispatcher* junto con la información que ya tenía recogida, los almacene en la base de datos.

6.5 Fuentes textuales: Medios tradicionales

6.5.1 Prensa digital

Con la llegada de la digitalización e Internet, la prensa tradicional ha evolucionado desarrollando portales Web donde plasmar su información periodística, dándole un toque más atractivo visualmente para el usuario. Este recurso es uno de los más utilizados por los usuarios, llegando a crecer al mismo ritmo que crecen los usuarios en Internet.

La prensa digital son portales Webs que se definen por su frecuencia de actualización. A diferencia de los diarios tradicionales, estos se pueden actualizar varias veces al día, pero en algunos este tiempo de actualización se ve aumentado a varios días, semanas o trimestres.

Los contenidos de las noticias no solo aportan texto, sino que pueden ser complementados con imágenes y vídeos, aportando un mejor detalle sobre la noticia publicada.

Los periódicos digitales suelen utilizar una plataforma como gestor de contenidos, donde la estructura está compuesta por secciones y los artículos de una misma temática son agrupados, además cuenta con diversos tipos de contenidos ofreciendo al lector la posibilidad de escoger entre diferentes temáticas.

Al igual que en los diarios tradicionales, cuentan con un criterio de cierre, donde se terminan los periodos de actualización. No obstante, normalmente se puede acceder a las ediciones pasadas desde la propia página.

A diferencia de la prensa tradicional este cuenta con una serie de características propias del medio:

- La información se puede recuperar de forma sencilla y rápida. A través de buscadores o los menús, se hace de forma fácil seleccionar la información importante.
- Es posible la distribución instantánea de los contenidos a un indeterminado número de usuarios.
- La posibilidad de edición, agregación y eliminación de contenidos, ofreciendo mayor flexibilidad que las publicaciones impresas.
- Permite que los usuarios interaccionen con el portal. Estos pueden dejar sus comentarios y opiniones.
- Los usuarios pueden personalizar el portal según sus intereses.

Así pues, muchos de los periódicos digitales cuentan con herramientas de interacción como *blogs*, los RSS, los comentarios, el registro de usuarios, la comunicación vía email, foros, chat…entre otros. Además, con el fin de ampliar la difusión de contenidos y facilitar el acceso a los usuarios, los periódicos digitales no solo se encuentran en su página Web, sino que a través de los canales sociales publican los sucesos o noticias más actuales. También, gracias a la aparición las herramientas de permuta de contenidos como YouTube o Flickr, la prensa digital tiene la capacidad de propagar su información a través de estos canales, facilitándole el acceso a los usuarios y pudiendo llegar a crear comunidades.

Cabe destacar que dentro de los periódicos digitales podemos encontrar dos vertientes según su naturaleza: los periódicos digitales con replica en Internet y los que son únicamente digitales. La diferencia entre ambos reside en que los primeros son periódicos que a parte de encontrarlos en su forma digital (Internet), también los podemos encontrar en su forma tradicional (papel), así pues, podemos hallar periódicos como *El País* o *El Mundo*, que tienen tirada en las dos vertientes. Sin embargo, también existen otros como *El Plural* que únicamente se encuentra en su formato digital.

La importancia de los periódicos digitales no solo reside en la elaboración de contenidos, sino en la gestión de información actualizada y contrastada en una jerarquía bien definida, distinguiendo entre lo significativo de lo irrelevante, y promoviendo la comunicación social.

Los periodistas han tenido que adaptarse al nuevo medio de comunicación, modificando muchos de los criterios habituales de su forma de trabajar. La veracidad de la información se ha visto contrarrestada por la velocidad de información, viéndose en algunos casos noticias que no han sido ciertas. Y esto viene ocasionado muchas veces por la gran cantidad de información que se transfiere por Internet.

Actualmente, los periódicos digitales reciben cantidades enormes de información por los usuarios, por gente especializada o simplemente por los periodistas buscando noticias en la red. Pero, esta información no siempre es verídica y debido a las ansias de ofrecer en primicia la noticia, muchas veces son publicadas sin ser comprobadas. Además, la capacidad de interconexión y la fluidez con la que circula la información por Internet, revela que la misma noticia puede ser encontrada en muchas otras páginas Web.

6.5.2 Ámbitos de clasificación

A diferencia de las otras fuentes de información, en este canal se puede conocer a priori las distintas páginas Web que son considerados como periódicos digitales, ya sea con replica online o únicamente digitales.

Así pues, a partir del censo de periódicos revisado por un experto se puede identificar de forma unívoca lo que es un periódico digital de lo que es otro tipo de fuente.

A diferencia de los *blogs* y los foros, los contenidos son desarrollados a partir de un equipo de redacción con unos roles bien definidos. Este equipo jerarquizado es el responsable de todo el portal, y se encargan de aportar nuevos contenidos, del mantenimiento y la distribución de estos, además definen cuando se realiza el cierre de la edición.

Intentan tener una comunicación casi en vivo donde los usuarios pueden comentar los contenidos pero en muchos casos sin una reciprocidad. A diferencia de los medios tradicionales, donde la información es de ayer y los comentarios son más meditados y reflexionados, en los digitales los comentarios son más desde lo personal, desde la pasión del usuario, dejando a un lado la racionalidad y la meditación.

Algunos de los periódicos digitales empelan la utilización de un foro, donde los usuarios puedan debatir o un *blog*, donde usuarios expertos hablen de un tema relacionado con una notica.

Así pues, los periódicos digitales son clasificados como generadores de noticias, que debido al censo cerrado que existe se pueden identificar con facilidad.

6.5.3 Particularidades de la recuperación

Como se ha podido observar en los puntos anteriores, los periódicos digitales son una fuente de información donde los contenidos pueden ser actualizados en diversas ocasiones durante el día, ofreciendo noticas de actualidad que en algunos casos pueden no estar contrastadas en su totalidad o influenciadas por los matices del poder.

Este canal tiene algunas limitaciones derivadas de todos los canales textuales implementados bajo las APIs de los buscadores y algunas ventajas propias de la fuente.

Igual como se vio anteriormente, uno de los inconvenientes deriva de la cantidad de información que es devuelta por las APIs utilizadas. Como máximo devuelven un fragmento muy reducido que en muchas ocasiones no es lo suficiente extenso como para interpretar correctamente el contenido de la información.

Otro, es la fecha de la información. Los *crawlers* de los buscadores asignan a la página Web una fecha la cual indica la última visita que hizo el *crawler*. Al utilizar las APIs de recuperación de información de los buscadores, e intentar buscar información sobre noticias actuales, se hace una tarea casi imposible, debido a que los resultados obtenidos vienen enmarcados por la fecha que propuso el *crawler* y no por la fecha de los contenidos de la página.

Debido a que al ámbito de actualización de este canal no es considerado continuo sino que se dilata bastante en el tiempo durante el día, se precisa de una monitorización con disparadores que comuniquen al sistema de recuperación de que se ha producido una nueva actualización en el canal.

No obstante, uno de las ventajas que aporta este tipo de fuentes es el censo cerrado de estos medios, es decir, se puede conocer a priori todos los medios digitales de prensa que se pueden encontrar en Internet. Mediante páginas Web como *prensaescrita* se pueden obtener un listado de los periódicos digitales.

Por todas las limitaciones que subyacen de la utilización de las APIs para este canal y por el censo cerrado que se proporciona, la recuperación de información de esta fuente se resuelve mediante la utilización de *crawlers* específicos que sean capaces de acceder a la página, analizarla y extraer la información más relevante.

6.5.4 Recuperación: Información y limitaciones

Para la recuperación de información en el canal de prensa digital se ha implementado un *crawler* específico que a partir de un listado de direcciones (semillas) Web a páginas de diarios online obtiene la información de página de forma sistemática y automática.

El *crawler* comienza visitando la primera URL especificada en la lista de periódicos y almacena su contenido. Este proceso es repetido hasta finalizar con la lista.

Una de las limitaciones que presenta es la obtención de información en el formato especificado por la página Web, haciéndose necesario la utilización de algoritmos que consigan extraer la información de lo que es código, como la identificación de títulos, la fecha y el contenido.

6.5.5 Implementación en Cosmos

Los periódicos digitales publican muchas noticias en la portada de sus páginas Web, por lo que el sistema ha de recuperar la información que únicamente tenga contenido relevante para la consulta del usuarios.

Por ello, la primera tarea una vez obtenido el contenido de la página Web por el *crawler*, es la de limpiar y organizar la información para identificar cuales son las partes de la estructura donde se ha publicado contenidos de noticias y descartar cuales no son relevantes para el usuario.

Otra tarea, es la de identificar la fecha correspondiente al contenido publicado. Para ello, una vez discriminado lo que no es relevante y estructurado los contenidos, se busca la fecha correspondiente a cada uno de ellos. Para identificarla, se hace uso de algunas técnicas de extracción de información.

Una vez estructurado el contenido de la página e identificado cuales son las noticias y cual es su información, el sistema almacena la información relevante para el usuario, en su lenguaje formal para posteriormente mostrárselo al usuario.

Toda esta extracción de información se basa en el estudio realizado por Chistian Kohlschütter [B.60].

En la siguiente figura se muestra las partes funcionales que se sigue en la obtención de información de estas fuentes.

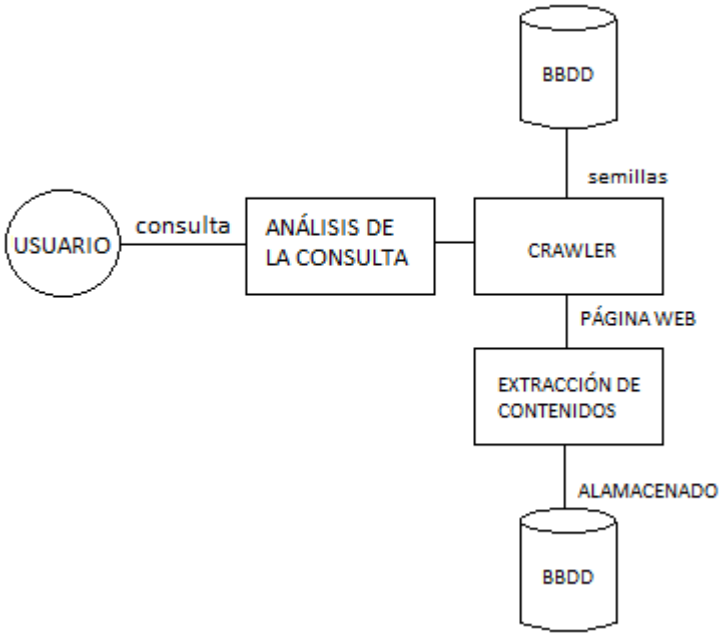

Figura 27. Proceso de recuperación en medios tradicionales

Al inicio, el usuario establece la consulta de información que será procesada para dársela al *crawler* específico. El *crawler*, obtendrá la lista de semillas de la base de datos, por la que irá recorriendo en busca de información que satisfaga a la consulta por el usuario. Una vez obtenida la página Web, se analiza y extrae el contenido de la información, quedando la información en un formato entendible por el usuario.

Al final esta información es almacenada en la base de datos para su posterior análisis. Al igual que en el anterior modulo de fuentes textuales se ha seguido una arquitectura paralelizada.

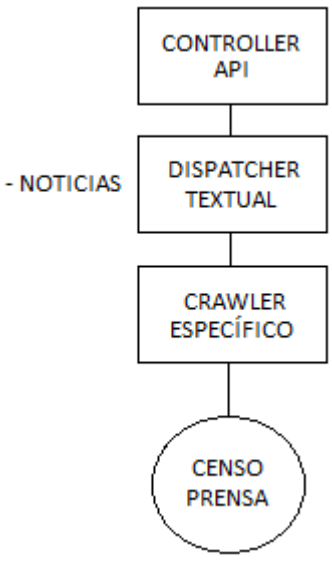

Figura 28. Arquitectura en medios tradicionales

El *controller API* se encarga de recoger todas las consultas que quedan por analizar. A partir de las consultas, estas son enviadas al *dispatcher textual* quien gestiona de forma paralela cada una de ellas. El *crawler* recibe la consulta y obtiene de la base de datos la lista de las semillas de las cuales tiene que consultar la información. En caso de encontrar coincidencia entre la consulta del usuario y la información de la página, este devuelve el contenido al *dispatcher textual*, el cual se encargará de filtrar y analizar la información dejándola en su forma textual para un mejor entendimiento por parte del usuario. Una vez acabado este proceso, la información es almacenada en la base de conocimiento.

6.6 Microblogging

6.6.1 Twitter

Twitter es una red gratuita donde los usuarios pueden aportar publicaciones de un máximo de 140 caracteres, llamados tweets. Como se puede apreciar, el tamaño de estos contenidos es de carácter reducido y se asemeja bastante al espacio utilizado en los mensajes de textos de los teléfonos móviles. Es por ello, que este tipo de canal es considerado como un sistema de mensajería instantánea.

Para consultar la información en esta red, no se hace necesario el registro de un usuario. De forma anónima, se pueden ir consultando los diferentes perfiles de los usuarios e ir captando la información de interés, siempre y cuando, el perfil del usuario consultado sea público. No obstante, si se precisa de la necesidad de escribir mensajes, organizarlos u otra tarea dentro de la red, es necesario registrarse y abrir una cuenta propia.

El servicio de Twitter cuenta con una mezcla de características de varios canales: los *blogs*, las redes sociales y la mensajería instantánea.

Este sistema tiene ciertas similitudes con los *blogs*, principalmente se relacionan por su sistema de publicación personal, la cronología de la información, los *micropost*, los enlaces permanente y las suscripciones RSS. No obstante, debido a su reducido tamaño en los espacios de los mensajes y a las características propias que adquiere de las redes sociales, este sistema se ha derivado en un servicio conocido como *microblogging*.

No obstante, este canal aporta sus características propias respecto a los otros. Deja a un lado la reciprocidad de las redes sociales, es decir, en Twitter, al ser un espacio público los mensajes se pueden consultar sin necesidad de estar relacionado con el usuario. Sin embargo, ofrece la posibilidad de seguir a un usuario en concreto, facilitando y organizando la lectura de los mensajes. Así pues, Twitter se convierte en un panel de lectura personalizado donde irán surgiendo mensajes de los seguidores que ha especificado el usuario. De igual modo, los usuarios que estén interesados en nuestras publicaciones pueden seguirnos para suscribirse a nuestro flujo de información. En cualquier caso, no es necesaria una relación simétrica. A diferencia de la mensajería instantánea, los mensajes pueden ser accedidos en un tiempo posterior al que fue publicado.

Quizá la característica o la propiedad más atractiva de Twitter es la informativa. Esta red se ha transformado en una de las primeras fuentes de información donde se hace público las últimas actualizaciones o novedades de los temas de interés. Su inmediatez y su continuo flujo de información generan grandes cantidades de información de actualidad donde los usuarios aportan sus opiniones referentes a un tema.

6.6.2 Ámbitos de clasificación

Si miramos hacía atrás en el tiempo, Twitter nace como un servicio sin un objetivo claro y algo confuso. La dificultad para definir en que consistía venía arraigada por la naturaleza de los mensajes de aquel momento: “Me voy a la playa”, “Estoy durmiendo”… Así pues, fue considerado como un servicio donde contar en cada

momento lo que le estaba sucediendo al usuario. Por esta peculiaridad, muchos usuarios no entendían cual era la finalidad y para que se podía utilizar, llegando a abandonarlo por completo.

Sin embargo, tras el paso del tiempo, han sido los propios usuarios los que le han dado un significado, una forma de ser, y es que, aunque Twitter no haya cambiado mucho desde sus inicios, ahora sí que tiene un objetivo claro, una utilidad para los usuarios y la información. Un objetivo mucho más definido e interesante, uno más atractivo y de interés, uno que constituye la clave del éxito de esta red: la información en tiempo real.

Twitter se ha ido convirtiendo en el canal de información más rápido del mundo. A través de sus tweets revela noticias en tiempo real, sucesos de actualidad y grandes cantidades de opiniones sobre temas actuales. Gracias a su facilidad y rápida publicación y a la limitación del texto del mensaje, los usuarios consiguen centralizar la información dejando al margen la argumentación y las florituras de otros canales de información.

El hecho de estar presente en todo el mundo y las diversas aplicaciones de uso que están surgiendo en torno a esta red, hacen que este servicio se haya convertido en uno de los canales de información con más auge y espontaneidad en el que las últimas noticias y sucesos se dan antes que en los medios tradicionales, permitiendo enriquecer el acto comunicativo mediante el conocimiento y la posibilidad de opinar en tiempo real.

Así pues, actualmente Twitter es considerado como la principal fuente de información en tiempo real en la que los usuarios dejan sus comentarios de los sucesos que se están produciendo actualmente, y su ámbito de clasificación es por tanto un medio acotado bajo un sistema de red social propio.

6.6.3 Particularidades de la recuperación

Como se ha visto anteriormente, Twitter es un canal de información en el que los contenidos se van agregando de forma continua y sin pausas, favoreciendo el discurso y el intercambio de opiniones. Debido a la aparición del periodismo urbano, a las cuentas oficiales de los principales medios de comunicación tradicionales y a las innumerables empresas que se encuentran dentro de esta red, se publican grandes cantidades de información de sucesos que están ocurriendo en el acto. Esta característica tiene algunas ventajas e inconvenientes frente a la recuperación de información.

La información en este canal se genera de forma rápida. Debido a su sencillez y a su limitación de caracteres, permite que un usuario escriba un contenido en un fragmento de tiempo reducido. Según el estudio [B.61] realizado por la herramienta Cosmos para la campaña de elecciones del 2011 en la noche del debate electoral, se recopilaron un total de 941.801 tweets publicados por 207.867 usuarios únicos. Estas cifras, confirman claramente la cantidad de información que se emite en Twitter, por lo que el sistema de recuperación en este canal ha de hacer frente a la capacidad de recoger la información en tiempo real y al almacenamiento de esta mediante técnicas de *big data*.

Con la aparición del fenómeno de la prensa urbana, los usuarios tienden a publicar los acontecimientos de una perspectiva de lo personal, así pues en muchas ocasiones los contenidos pueden ser malinterpretados y confundidos.

Una de las principales características de este medio es la peculiaridad de tener los perfiles abiertos. La filosofía de Twitter es la de compartir, debatir, opinar e intercambiar información, no obstante, permite a los usuarios el restringir sus perfiles al resto de usuarios.

Debido a las características propias de este canal, el sistema de recuperación de información ha de ser capaz de obtener los contenidos en tiempo real, ya que la información está en flujo continuo, y además, en algunos casos ha de conocer y filtrar la información que pueda ser fraudulenta o de poco interés para el usuario.

Por lo tanto, los dos principales hitos en la recuperación de información en el canal Twitter son el tiempo real y el *big data*.

6.6.4 Recuperación: Información y limitaciones

A través de la API Twitter se pueden realizar todas las operaciones como si estuviésemos en la Web, no obstante presenta una serie de limitaciones propias del servicio.

Este servicio, en realidad está compuesto por tres APIs diferentes. El *StreamingAPI*, *Search API* y *REST API*. Los dos primeros son los implementados para el sistema de recuperación de información del proyecto, y a continuación se presenta una breve descripción de su utilización y cuales son sus peculiaridades.

El *Search API* ofrece la recuperación de tweets como máximo de 7 días atrás desde el momento que se ejecuta la consulta o bien los últimos 1500 tweets. Las consultas de información se pueden definir mediante una palabra, un usuario (@), un *hashtag* (#) o por la combinación de algún tipo de operador. Es posible la utilización de filtros por idioma, localización o usuarios, y los datos son presentados en JSON o Atom.

No se requiere autentificación por parte del usuario, Twitter registrará la IP desde donde se hace la consulta para ir controlando la cantidad de peticiones realizadas. No obstante, también permite el *loging* de usuario a la hora de realizar las peticiones de consulta, controlando de esta forma el acceso a la API a través de la información proporcionada de autentificación del usuario por el sistema de recuperación de información. Para esta segunda modalidad, se hace necesario dar de alta un usuario en Twitter, y registrar una aplicación de desarrollador en Twitter App. De esta forma, se obtiene cuatro claves de acceso: el *consumer key*, el *consumer secret*, el *acces token* y el *acces secret*. Todas ellas, son necesarias para que el sistema pueda identificarse en Twitter mediante el protocolo de autentificación OAuth.

El número de peticiones está limitado a 150 a la hora por usuario o por IP para los usuarios no autentificados. Este número de peticiones se ve incrementado a 350 a la hora cuando el usuario se autentifica en Twitter. Sin embargo, para ambos métodos se debe esperar un par de segundos entre una petición u otra, sino el sistema puede que corte la transferencia o devuelva resultados erróneos. Una vez pasado 24 horas todos los límites suelen ser revocados.

La información obtenida a través de este modelo se muestra en la siguiente tabla.

Resultado	**Descripción**
Tweet	Información en texto expresada por un usuario.
id_tweet	Identificador del tweet.
id_user	Identificador del usuario que escribe el tweet.
id_retweet	Identificador del usuario al que se le hace retweet. Este valor únicamente existe si el tweet que recogemos es un retweet de un tweet.
Lang	Idioma en el que está el tweet.
Longitude	Indica la longitud de donde proviene el tweet. Este valor solo estará disponible si el usuario a geolocalizado el tweet.
latitude	Indica la latitud de donde proviene el tweet. Este valor solo

	estará disponible si el usuario a geolocalizado el tweet.
Date	Indica la fecha en que se publicó el tweet.

Figura 29. Twitter: Tabla de resultado

Como se puede apreciar da información del *tweet*, pero únicamente el id del usuario. Esto obliga a realizar una segunda petición para obtener los datos actualizados del usuario. Este problema en la medida de lo posible lo resolvemos almacenando la información del usuario en unas tablas auxiliares, evitando así tener que realizar nuevas peticiones para obtener la información del usuario en concreto. No obstante, somos conscientes de la variabilidad que existe en Twitter, y que la información del usuario se va actualizando constantemente. Es por ello, que se precisa de un método auxiliar, que cada cierto tiempo vaya actualizando los perfiles de usuarios más activos en la tabla de la base de datos. Esta posibilidad presenta una serie de ventajas e inconvenientes: las peticiones se ven reducidas, por lo que se puede hacer más peticiones de recuperación de información de tweets, además, de esta forma se consigue evitar el tener que obtener la información actualizada de todos los usuarios que hayan escrito un *tweet*, evitando así los usuarios menos activos y centrándose únicamente en los más participativos. Por el contrario, cabe la posibilidad de que un usuario no tenga la información de su perfil actualizada en el sistema.

En el *Streaming API* se realiza una conexión HTTP permanente a los servidores de Twitter. A través de esta conexión se recibe un flujo continuo de información prácticamente al mismo tiempo que son enviados los *tweets*. Es necesario estar autentificado mediante el protocolo de autentificación *OAuth*.

Esta modalidad permite diferentes métodos para obtener la información:

- *Links*: obtiene la cantidad de tweets públicos que contienen enlaces.
- *Filter*: obtiene la cantidad de tweets acotados por uno o más filtros.
- *Firehose*: obtiene todos los tweets públicos.
- *Sample*: devuelve de forma aleatoria los estados públicos.
- *User*: obtiene toda la información referente a un usuario (actualizaciones, información…). Se puede seguir hasta 1000 usuarios a la vez.
- *Retweets*: obtiene los *retweets* que se producen en Twitter.

A continuación se muestra los datos que se pueden obtener a través de esta modalidad.

Resultado	**Descripción**
Tweet	Información en texto expresada por un usuario.
Name	Nombre del usuario que escribe el tweet.
id_user	Identificador del usuario que escribe el tweet.
id_retweet	Identificador del usuario al que se le hace retweet. Este valor únicamente existe si el tweet que

	recogemos es un retweet de un tweet.
Retweeted	Identifica si el tweet es un retweet.
Date	Indica la fecha en que se publicó el tweet.
Hashtags	Conjunto de etiquetas que aparecen en el tweet.
Urls	Conjunto de direcciones que aparecen en el tweet.
user_mentions	Conjunto de usuario que aparecen en el tweet.
coordinates	Coordenadas de geolocalización del tweet.
user_name	Nombre del usuario que ha escrito el tweet.
profile_image	Imagen del usuario que ha escrito el tweet.
Location	Localidad registrada por el usuario.
time_zone	Zona horaria del usuario.
Protected	Identifica si el perfil del usuario está protegido.
followers_count	Cantidad de seguidores del usuario que ha escrito el tweet.
friends_count	Cantidad de usuarios a los que sigue.
listed_count	Cantidad de listas que tiene el usuario.

A pesar de las ventajas que proporciona esta modalidad de API, como la obtención de información en tiempo real, la posibilidad de conseguir más información y estar actualizada, la sencillez con la que se realiza la conexión a la API, la posibilidad de seguir la información filtrada por usuarios, localidades, idiomas o temáticas, también, presenta una serie de limitaciones que hay que considerar.

Quizá la mayor limitación que presenta es que únicamente devolverá el 1% de los tweets que no excedan del *Firehose* de Twitter. Es decir, si en un momento concreto los tweets que se quieran recuperar superan al 1% del total de tweets publicados en Twitter, la cantidad de resultados estará limitada por ese porcentaje.

La velocidad del flujo de información vendrá determinada por el ancho de banda disponible en cada momento y la sobrecarga de los servicios de Twitter.

Las conexiones a los servidores se pueden romper, por lo que se precisa de un sistema que recupere la conexión y vuelva a ejecutar el seguimiento de la información. Si el sistema intenta realizar la conexión repetidamente o se demora en la conexión, Twitter bloqueará la conexión y penalizará al usuario. Además, al tener la necesidad de estar autentificado por usuario, no es posible la ejecución de dos *streamming* al mismo tiempo en diferentes sistemas. En caso de ejecutarlo en varios sistemas, Twitter repartirá la información enviada entre las diferentes ejecuciones. No obstante, si el cliente de *streamming* deja de leer datos de Twitter durante un cierto periodo de tiempo, Twitter cerrará la conexión.

No obstante para cualquier API de Twitter existe una lista blanca (*White List*) que permite hacer unas 20000 peticiones a la hora. Esta estimación de peticiones fue obtenida por los registros de publicación de Twitter, los cuales estiman que con esa cantidad es suficiente como para obtener todos los tweets. Eso sí, la integración en este tipo de lista es muy complicado y la mayoría no pueden acceder, además, de que Twitter anunció que este servicio pronto sería eliminado para evitar así la saturación del sistema.

6.6.5 Implementación en Cosmos

Para la recuperación de información de este canal se ha realizado la implementación de dos de los tres tipos de APIs ofrecidas por Twitter.

El primer módulo utiliza la API Search de Twitter, que mediante ella se van obteniendo los resultados especificados por la consulta de información del usuario de forma paralela para cada proyecto y en la medida de lo posible, en tiempo real.

Como se ha podido observar anteriormente, este sistema de la API de Twitter no devuelve la información del usuario, por lo que se precisa de una tabla donde se esté almacenando la información del usuario.

Cuando se obtiene un tweet, se identifica el id del usuario que por el cual se hace una consulta a la base de datos en busca de la información del usuario. En caso de no existir, se realiza una petición a la API de *User* de Twitter para obtener la información del usuario, y almacenarla en la base de datos.

Cabe destacar, que existen procesos adicionales que van realizando peticiones a la API de *User* de Twitter para ir actualizando la información de los usuarios más influyentes.

Así pues, con el conjunto de estos dos módulos se obtiene la información de este canal y además las características de los perfiles de los usuarios.

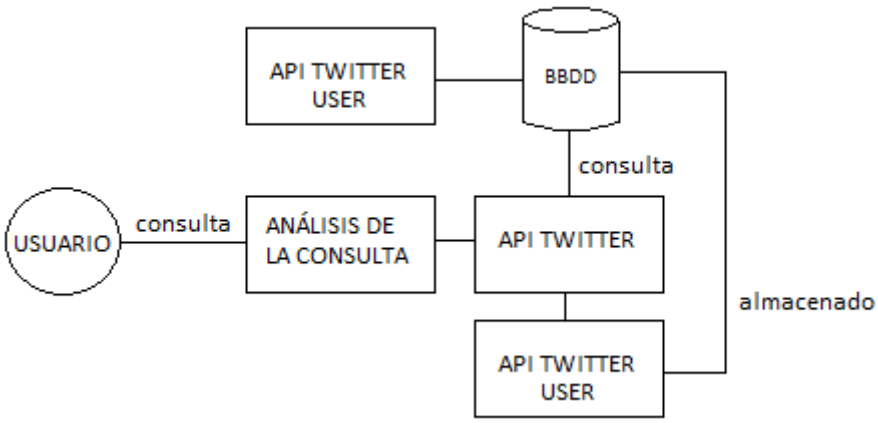

Figura 30. Proceso de recuperación en Twitter

En la siguiente imagen se puede ver la arquitectura seguida para el canal de Twitter.

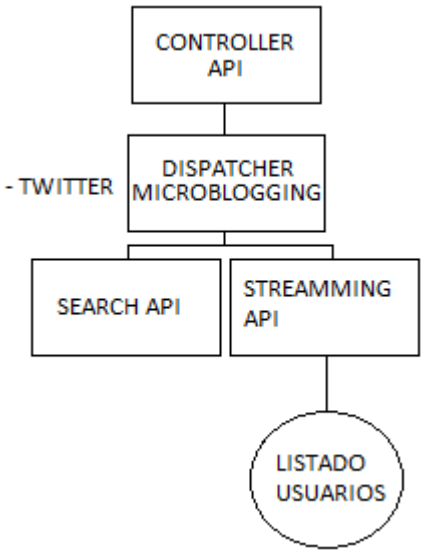

Figura 31. Arquitectura en Twitter

El *controller* API recibe todas las consultas por parte de los usuarios. Estas consultas son enviadas al dispatcher que se encargará de recoger toda la información posible. Una vez recogida la información, se comprueba si el usuario existe en el censo de usuarios, en caso de que no exista, se utiliza la API User de Twitter para obtener su información y posteriormente almacenarla. Una vez realizado esto, toda la información es almacenada en la base de conocimiento. Debido a que es una recuperación en tiempo real el controller esta continuamente recuperando las consultas de los usuarios y ejecutando tantas veces como sea necesario el *dispatcher*.

6.7 Redes sociales

6.7.1 Facebook

Facebook es una red social que permite conectar a diferentes usuarios de forma gratuita. Este servicio ofrece a los usuarios un espacio personal donde se puede compartir tanto información multimedia como textual entre los distintos usuarios, además de participar en eventos y otros servicios.

Debido a su facilidad de uso y a la oferta de servicios que proporciona fomentando la interactividad del usuario, han hecho que esta red se haya convertido en poco tiempo en una de las más utilizadas y más grandes del mundo.

La peculiaridad de esta red social reside en los enlaces entre los usuarios. Mediante la aceptación de amistad por parte de ambos usuarios, estos pueden compartir todo tipo de información, es decir, con este vínculo un usuario puede ver lo que el otro usuario publique. No obstante, existen reglas que pueden bloquear el acceso a algunos contenidos. De la misma forma, un usuario que no acepte la amistad no podrá ver los contenidos del otro, a no ser que este tenga un perfil completamente público.

Además, otra de las características de esta red social, es el botón *me gusta*. Este botón es utilizado por los usuarios para valorar el contenido o no. Es decir, a un usuario que le haya gustado el contenido del comentario o *post*, puede dejar un *me gusta* expresando su conformidad con el mensaje, y además añadir un nuevo comentario.

6.7.2 Ámbitos de clasificación

La filosofía de las redes sociales trasciende de los foros tradicionales. En estos, un usuario intercambia información y opiniones sobre un tema en particular en una comunidad cerrada. Este intercambio se hace de forma gradual y puede dilatarse a lo largo del tiempo, dependiendo estrechamente de la actividad por parte de los usuarios. Aunque este tipo de canal sigue en funcionamiento, se ha visto apartado debido a la aparición de las plataformas de redes sociales.

Quizá una de las principales diferencias es la gran variedad de servicios que aportan las redes sociales. Esto provoca que los usuarios interaccionen más con los contenidos y fomenten la creación de círculos sociales de intercambio libre y continuo de información rica en opiniones. Además, en las redes sociales no se cuenta con un moderador para administrar los contenidos, sino que es el propio usuario el que gestiona toda la información publicada.

Las redes sociales, pretenden alcanzar el flujo de información en tiempo real igual que en los *microblogging*. Estar informado continuamente y de forma inmediata supone una cualidad que los foros no pueden ofrecer. En la mayoría de ocasiones, este flujo de información se hace desde un perfil de usuario en concreto, el cual contiene un información de perfil y una actividad asociada, por lo que, el anonimato de los foros se ve arraigado en este ámbito.

Otra cualidad de estos sistemas son los perfiles de usuarios y, la información presente y compartida. A diferencia de los foros, en este tipo de herramientas los usuarios cuentan con un perfil mucho más completo e interactivo, donde es posible agregar ficheros multimedia y hacer uso de otras herramientas que conformen la identidad digital del usuario. El intercambio de información, de elementos multimedia y de archivos son puntos clave frente a los tradicionales foros.

6.7.3 Particularidades de la recuperación

Facebook ha cambiado durante los años, de pasar a ser una red social enfocada a la producción de información del estado de un usuario, orientada principalmente a los personas y en especial a los jóvenes, a una perspectiva más allá del intercambio de contenidos personales, donde los medios de comunicación, las empresas y las organizaciones lo utilizan para aumentar su propagación de información.

Esta red social permite tres modalidades de registro, y van ligados estrechamente con los ámbitos en los que se va utilizar:

- La individual, enfocada directamente a los perfiles personales de los usuarios.

- La institucional o páginas, las cuales están pensadas para organizaciones, permitiéndoles tener un sitio Web público fuera de los límites de Facebook. Además, los usuarios pueden inscribirse como seguidores sin la necesidad de la autorización de la organización, permitiéndoles así un flujo continuo de información.

- La de grupos, que permite la congregación de una comunidad con un tema en común. Entre sus características más importantes cabe destacar la facilidad de creación y promoción de este tipo de páginas, así como una mejor visibilidad del grupo.

Si centramos la atención en el ámbito personal, los usuarios pueden intercambiar contenidos y elementos multimedia entre sus amigos. Su finalidad reside en el mero entretenimiento, ya sea bien utilizando algún servicio que ofrezca la red social, compartiendo información con el resto de usuarios, etiquetando contenido multimedia…en resumen, dentro de una red social un usuario tiene infinidad de cosas con las que entretenerse e ir pasando el tiempo.

Por otra parte, si nos centramos en el ámbito empresarial, sin ninguna duda la aparición de este tipo de herramientas ha favorecido en un cambio espacial y temporal debido al efecto globalización que presenta. Las empresas, pueden darse a conocer a nivel mundial a través de este canal de información, y así captar un mayor número de espectadores, que a través de otro medio posiblemente no sería factible.

Por último, desde el punto de vista político, la posibilidad de participación ciudadana y acercamiento a estos son unos de los punto clave en la utilización de las redes sociales en este ámbito. Así pues, han funcionado como lugares de reflexión e intercambio de ideas entre los usuarios y los políticos.

Así pues, este canal se está convirtiendo en una de las fuentes de información donde los medios de comunicación y las organizaciones publican todas las novedades destacables que consideren oportunas.

Al contrario de las páginas individuales, el resto se comportan casi igual como una página Web, donde la diferencia radica quizá en la expansión de la información a nuevos oyentes y la alta posibilidad de participación que ofrecen.

No obstante, la comunicación entre usuarios no es igual como en los *microblogging*, aquí trasciende más en el tiempo, además de que los comentarios no son de carácter reducidos, sino que debido a la naturaleza de la que proviene, la información se pública en forma de *post*. La filosofía de Facebook es la de conectar usuarios, con la finalidad de que puedan ver sus aportaciones, mientras que la idea de los *microblogging* es dejar al usuario la libertad de recibir la información interesante para él sin tener la obligación de seguir a nadie.

Además, el centro de la información de interés prácticamente radica en torno a las páginas y los grupos, a no ser que se quiera obtener información concreta de un usuario especifico.

En las redes sociales se promueve la compartición de información en forma de comentarios, vídeos, fotografías…además de contar con servicios adicionales de convocatoria como los eventos, los calendarios…Su carácter es mucho más privado, la comunicación es entre el circulo social del usuario. A diferencia de otros medios, se da más importancia a la relación entre las personas que a la información que se está generando sobre un tema en concreto.

Así pues, este tipo de canal presenta una serie de ventajas e inconvenientes en cuanto a la recuperación de información.

La información se hace prácticamente continua, pero como hemos visto anteriormente, la mayoría de información que se genera es de tipo personal y sin poca relevancia. Centrados en las páginas y los grupos, la información es constante, pero ya en un marco temporal algo más extenso. Es por ello, que no se precisa de una constante captación de información.

Debido a privacidad de algunos perfiles de usuarios se hace imposible la recuperación de información.

La información aportada por la API en muchas ocasiones es bastante pobre e incompleta para hacer un buen análisis.

6.7.4 Recuperación: Información y limitaciones

Por confidencialidad con la aplicación de escucha activa de Cosmos, la recuperación de información utilizada en este tipo de canal queda reservada por derechos profesionales.

La recuperación en este canal se basa en el siguiente modelo de objetos, relaciones y búsquedas.

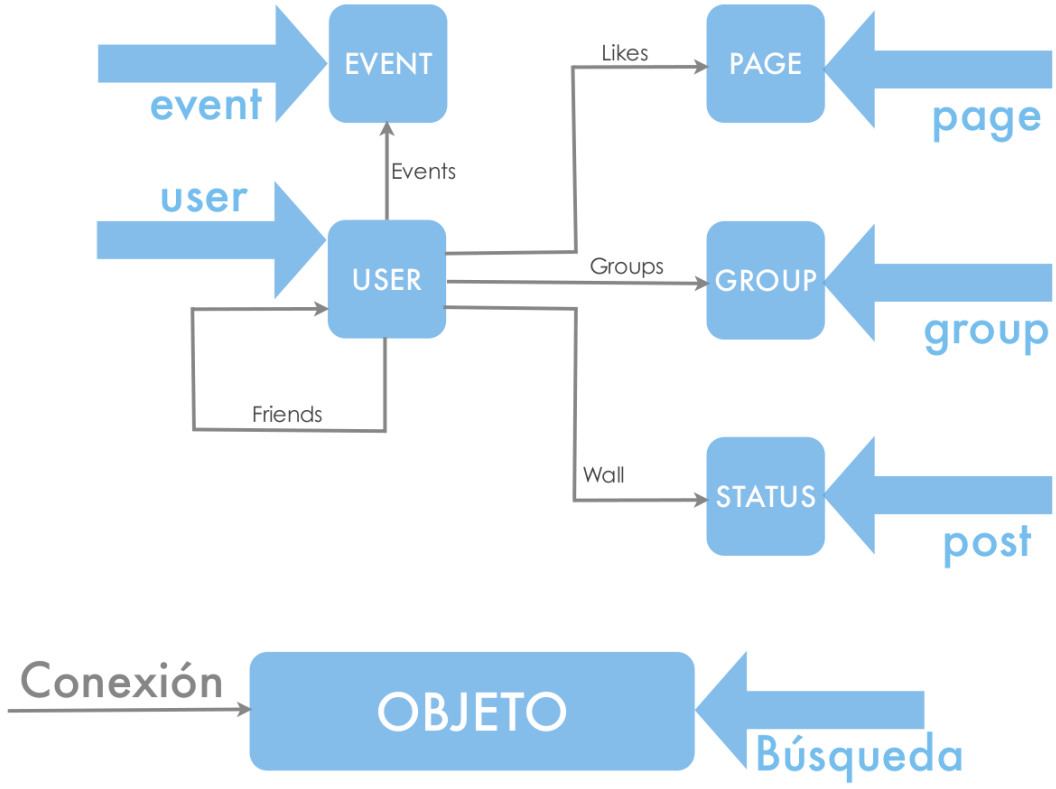

Figura 32. Modelo de objetos Facebook

6.7.5 Implementación en Cosmos

No obstante, se va ha describir sin entrar en detalles y luego se verá una ilustración donde se puede apreciar la arquitectura del módulo.

La recuperación de información en esta red social no es una tarea sencilla. La información que se presenta mediante la utilización de su API, en muchas ocasiones resulta pobre y de poco valor. Esto es debido a las continuas restricciones de seguridad que se establecen en la red social. Así pues se precisa de un módulo que pueda obtener mayor información al usuario, ya sea a partir de esta vía u otra.

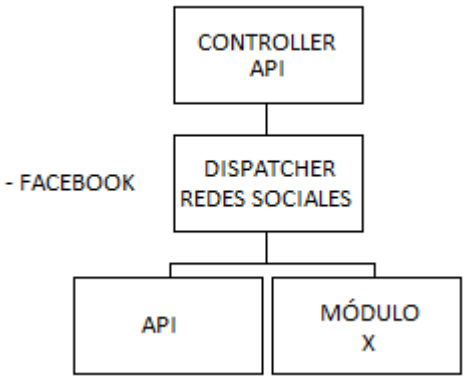

Figura 33. Arquitectura en Facebook

6.8 Fuentes multimedia: YouTube

6.8.1 YouTube

YouTube, al igual que otras herramientas de intercambio de contenidos ofrece un servicio gratuito donde intercambiar vídeos que son aportados generalmente por usuarios, aunque también es muy utilizado por las empresas y organizaciones para hacer campañas de *marketing online* y difundir así su mensaje llegando a un número mayor de espectadores. No obstante, su mayor peculiaridad ha sido el impacto social que ha suscitado en los usuarios, invitándoles y animándoles a generar vídeos personales y luego compartirlos en el canal.

La filosofía de YouTube es la de compartir información multimedia de tipo video. Los usuarios, a través de un registro previo, pueden subir y bajar videos, compartirlos, valorarlos y clasificarlos, enviar comentarios, mensajes personalizados, subscribirse a otros usuarios, publicarlos en otras páginas Web…Además, permite que los usuarios se puedan relacionar a partir de las suscripciones al canal del usuario. Como se puede apreciar, se ofrecen grandes cantidades de servicios al usuario.

Tan grande ha sido el impacto en la sociedad, que los usuarios se han vuelto reporteros y creadores de videos y en muchas ocasiones captan movimientos sociales que a través de otros medios o canales de comunicación no son publicados.

Descartando el mero hecho de que este canal es utilizado por algunos usuarios con el único fin de entretenimiento, YouTube es considerado como un canal en el que las contribuciones por los usuarios convergen con el periodismo profesional, un canal de protesta, un canal de información sin (casi) barreras…un canal libre de expresión mediante videos.

Una imagen vale más que mil palabras, pero un vídeo vale más que mil imágenes.

6.8.2 Ámbitos de clasificación

Llegados a este punto, hay que diferenciar entre lo que son redes sociales y servicios de red social. En los primeros se centran en ser estructuras sociales establecidas por nodos de unión interdependientes, como Facebook, mientras que los segundos son comunidades de usuarios que comparten intereses en común, como YouTube o Flickr, como veremos a continuación.

Se puede pensar que por el hecho de registro, el de poder compartir, el de participar y comentar, tener un perfil y modificarlo...YouTube o Flickr sean una red social, pero no es así. Está claro que permite muchas de las funcionalidades que podemos encontrar en una red social, pero este tiene un objetivo claro, el de que los usuarios compartan. El hecho de que permita relacionarse unos usuarios con otros, queda en un segundo plano bastante alejado del fin que persigue.

YouTube, no es un canal de los medios de comunicación televisivos, sino una página donde se brinda la oportunidad a los usuarios de compartir y participar en el intercambio de contenidos multimedia, y de opinar e intercambiar ideas. Si es cierto, que muchos de los usuarios consideran YouTube como un sustituto de la televisión.

YouTube ofrece una gran posibilidad de comunicación y *marketing online* para las empresas y organizaciones en el que a través de vídeos anima a los usuarios a participar en la aportación de opiniones y comentarios sobre los servicios o actividades que presentan.

Igual que ocurre en los foros, YouTube presenta unas condiciones y normas de uso muy estrictas, basadas en el fundamento de la responsabilidad y el respeto, además que los temas son de gran diversidad.

Así que, este canal de información aporta contenidos en formato video a la escucha activa. Aunque, la cantidad de videos y las visitas aumenten por minutos, la información aportada es muy personal y enfocada al ocio, por lo que no se puede considerar como un canal de información relevante continua.

6.8.3 Particularidades de la recuperación

El proceso de recuperación de información en este canal, se hace en algunos aspectos algo tedioso e incoherente. La dificultad con la que se presenta la búsqueda de información multimedia por los buscadores convencionales, hace que se precise de herramientas un poco más sofisticadas que entiendas el lenguaje multimedia. No obstante, en muchas ocasiones esto no está en práctica y se utilizan buscadores textuales que utilizan las descripciones y las etiquetas proporcionadas a los contenidos multimedia para ser comparados con la consulta por parte del usuario.

Así pues, para ayudar en la búsqueda de información multimedia en estos canales se precisa de que la información almacenada este correctamente etiquetas y categorizada. Sin embargo, esto presenta una serie de inconvenientes centrados en el usuario:

- Al dejar al usuario realizar este tipo de acciones, en muchas ocasiones la categorización o el etiquetado no son los adecuados, llegando a confundir los resultados obtenidos.

- Otro aspecto a tener en cuenta, es la descripción que se le asigna al elemento multimedia. Los usuarios pueden designar una descripción, que en muchas ocasiones no se parece a lo que el elemento multimedia quiere transmitir.

- Sin embargo, otro punto a tener en cuenta es la posibilidad de desconocimiento del origen del elemento multimedia, es decir, si no se conoce de en que consiste o que pertenece, muy difícil se hará el poder etiquetarlo.

6.8.4 Recuperación: Información y limitaciones

Para utilizar la interfaz [E.31] se ha utilizado la librería java *gdata-youtube-1.0.jar*, que a partir de su integración en un proyecto Java, facilita el acceso y consumo a la API de YouTube.

No obstante, la API al igual que todas las vistas hasta ahora, necesita de una clave para poder hacer uso de la interfaz. Debido a que YouTube es un producto de la empresa Google, para obtener la clave de acceso, se realiza con los mismos pasos que con la API de Google.

En cuanto a las restricciones, la API no presenta un límite total de peticiones, sin embargo, si que tiene restricciones en el tiempo entre las peticiones y la cantidad de videos devueltos. Llegados a pedir demasiadas consultas en un tiempo muy breve, la interfaz devuelve un mensaje de error. Una cualidad que ofrece esta interfaz, es que no penaliza por sobrepasar las peticiones, sino que únicamente se deberá de esperar un tiempo reducido pero prudente antes de realizar otra petición de consulta. Por otra parte, la máxima cantidad de vídeos esta impuesta en 999 resultados.

Los resultados pueden ser mostrados tanto en JSON como en XML. La información obtenida se puede ver en la siguiente tabla.

Resultado	**Descripción**
Id	Id del vídeo.
Uploaded	Fecha en la que se subió el vídeo.
Updated	Fecha en la que se modificó el vídeo.
Uploader	Nombre del usuario que subió el vídeo.
Category	Categoría seleccionada por el usuario.
Title	Titulo especificado por el usuario.
description	Descripción especificada por el usuario.
Tags	Listado de etiquetas que engloban el vídeo.
thumbnail	Listado de imágenes en formato disminuido que representan porciones del vídeo.
Duration	Tiempo que dura el vídeo.
likeCount	Número de usuarios a los que les gusta el vídeo.
ratingCount	Número
url	Dirección Web del vídeo.
comments	Listado de los comentarios del vídeo.
Iduser	Id del usuario quien ha subido el vídeo.

Figura 34. YouTube: Tabla de resultado

6.8.5 Implementación en Cosmos

Para la implementación en la herramienta de Cosmos se ha utilizado la interfaz de programación de YouTube.

En la siguiente ilustración se puede ver las fases que contempla la recuperación en este canal.

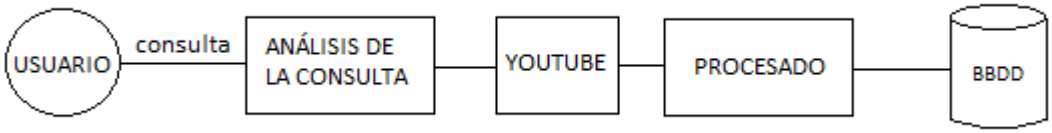

Figura 35. Proceso de recuperación en YouTube

El usuario realiza una consulta de información que es analizada y procesada para enviársela a la API de YouTube. Una vez analizada es enviada a la interfaz de programación quien se encargará de devolver los resultados de YouTube. Estos resultados son procesados para su posterior almacenamiento de la base de conocimiento.

Al igual que el resto de canales, en este se hace necesaria la implementación paralelizada, por lo que la arquitectura general del canal quedaría de la siguiente forma.

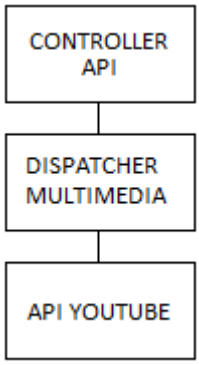

Figura 36. Arquitectura en YouTube

Donde el *controller* API obtiene todas las consultas por parte de los usuarios, y son trasladadas al *dispatcher* multimedia quien las procesará. Una vez procesadas, son enviadas a la API de YouTube la cual devolverá los resultados al *dispatcher*, para que este los gestione y los almacene en la base de datos.

6.9 Fuentes multimedia: Flickr

6.9.1 Flickr

Al igual que YouTube, es una herramienta de intercambio de contenidos, pero en este caso sirve para compartir y almacenar elementos multimedia de tipo fotografía, bajo una serie de condiciones de usabilidad que favorecen a la buena gestión de los contenidos. Además, en las últimas versiones permite la subida y compartición de vídeos por parte de los usuarios, pero esta no es la filosofía por la que nació.

Quizá sea la mejor herramienta online para la gestión, visualización y compartición de fotografías. Se caracteriza por ser una comunidad de usuarios en los que comparten aficiones entorno a la fotografía. Los usuarios tienen la capacidad de administrar las imágenes pudiendo etiquetarlas y comentarlas.

Este servicio permite que una comunidad de usuarios, compartan fotografías desde cualquier parte del mundo, con el objetivo de intercambiar impresiones para la realimentación de conocimiento del usuario.

Desde su aparición, no solo ha sido utilizado por los usuarios, sino que debido a la gran cantidad de imágenes que se pueden encontrar en los diarios digitales, esta herramienta se ha convertido en un fuerte aliado para ellos. El almacenamiento de miles de millones de imágenes se puede externalizar a este canal.

Con esto y las aportaciones de los usuarios, quizá sea una de las mayores fuentes de información de imágenes conocidas hasta el momento, donde no solo consiste en aportar, sino que también se puede buscar, geolocalizar, recuperarlas en una página Web y extraer información de contenido informativo.

No obstante, el uso de esta herramienta desde la empresa o las organizaciones se encuentra aún en usa forma muy retraída. Quizá la mayor restricción por la que llega a no ser usado desde este ámbito, es la de que no puede ser utilizado para fines comerciales. Sin embargo, con algo de creatividad esta restricción deja de ser una barrera para las empresas y organizaciones. Así pues, es utilizado para una mayor difusión de la información entorno a los eventos, catálogos, actividades…

6.9.2 Ámbitos de clasificación

En Internet podemos encontrar infinidad de imágenes digitales dispersas por el ciberespacio. El auge de la fotografía digital actualmente y las herramientas de intercambio de contenidos han favorecido a la generación de herramientas como Flickr.

Los inicios de Flickr provienen de un juego multiusuario, así pues no es de extrañar que una de las características que sostiene esta aplicación a parte de su facilidad de uso, sea la capacidad de participación por parte de los usuarios.

Una de las características potenciales de esta herramienta es la folksonomía. Flickr fue una de las pioneras en integrar la posibilidad de etiquetado social. Así pues, los usuarios pueden integrar descriptores a las imágenes en base a sus conocimientos y perspectivas personales. Este etiquetado es una técnica altamente escalable, moldeable y con un coste casi nulo para los administradores.

Así pues, este canal se puede clasificar como el sistema de información fotográfica por la que a través de su sistema de etiquetado permite obtener información en formato de imagen para la escucha activa.

6.9.3 Particularidades de la recuperación

Como se ha visto anteriormente, uno de los problemas de Flickr es la libertad con la que se presta al usuario el etiquetado de las fotografías, resultando en algunas ocasiones unas etiquetas poco precisas y llenas de ruido. No es para nada un sistema de calidad en el etiquetado y la comparación de imágenes fotográficas, su mecanismo se centra en la limpieza y el bajo coste, acorde a la actual caótica Web social. No obstante, los usuarios hacen una gran labor por intentar construir un sistema de calidad basado en los descriptores de las fotografías, fundamentado en la recuperación de información y en la compartición de esta.

Debido a la posibilidad de organización de las imágenes en función de la geografía, el sistema ofrece una recuperación de información centrada en un lugar concreto.

Otra cualidad de recuperación es la temática. La posibilidad de que los usuarios introduzcan los temas a los que pertenece su fotografía, puede afectar a la veracidad de los resultados obtenidos. Las inimaginables clasificaciones que puede un usuario aportar a una fotografía desbordan en algunos casos la calidad de información.

A través de la minería de datos de Flickr, este presenta un mecanismo por el cual las etiquetas de los usuarios son agrupadas según la concurrencia de estas, permitiendo la generación de *clusters* que agrupen una serie de imágenes con un mismo sentido.

6.9.4 Recuperación: Información y limitaciones

Para el consumo de la API de Flickr [E.32] se ha desarrollado una aplicación que interaccione con la interfaz de programación de Flickr. Esta interfaz se presenta para los desarrolladores que la vayan a usar con fines no comerciales, sin embargo, se puede llegar a un acuerdo previo para poderse utilizar dentro de este ámbito.

Para ello, se hace necesario el registro de una cuenta de correo de Yahoo! mediante el cual, se registra una aplicación para desarrollador para obtener la clave (*APIKey*).

En cuanto a las limitaciones que presenta esta interfaz, el máximo de resultados obtenidos esta establecido en 500 fotografías. No obstante, no presenta restricciones en cuanto al número de peticiones realizadas.

Los resultados obtenidos se presentan en formato JSON y son lo siguientes.

Resultado	**Descripción**
id	Id de la fotografía.
url	Dirección Web de la fotografía.
smalsquareurl	Fotografía en tamaño reducido.
mediumurl	Fotografía en tamaño medio.
largeurl	Fotografía en tamaño grande.
title	Titulo especificado por el usuario.
description	Descripción especificada por el usuario.
dateposted	Fecha en la que subió la imagen.
datetaken	Fecha en la que se tomó la fotografía.
lastupdate	Última actualización de la fotografía.
comments	Lista de los comentarios de la fotografía.
userid	Id del usuario quien ha subido la fotografía.
farm	Conjunto al que pertenece dentro de Flickr.

Figura 37. Flickr: Tabla de resultado

6.9.5 Implementación en Cosmos

Para esta implementación se procede igual que como en la de YouTube.

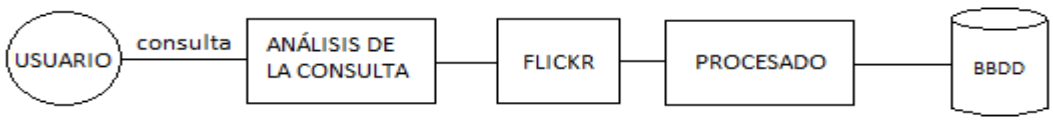

Figura 38. Proceso de recuperación en Flickr

El usuario realiza una petición de información al sistema. El módulo de análisis de consulta se encarga de interpretar la consulta de información, para posteriormente transferírsela a la API de Flickr. Los resultados son procesados y almacenados en la base de datos para su posterior análisis.

En cuanto a la arquitectura seguida, es la misma que en el anterior canal multimedia.

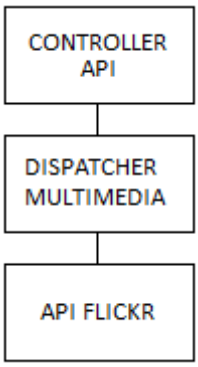

Figura 39. Arquitectura en Flickr

6.10 Una arquitectura de recuperación en tiempo real

En Internet los ciudadanos hablan, y lo hacen de tantas cosas que sólo es posible atenderlas si realizamos una efectiva Escucha Activa de los asuntos que nos interesan, y especialmente de los que les interesa a los ciudadanos. Pero escuchar todo esto significa recuperar y procesar unas cantidades ingentes de información con una cantidad considerable de métodos capaces de extraer conocimiento de esos inmensos volúmenes de datos sin estructurar y presentarlos de manera inteligible, usable, innovadora, para el usuario. Todo ello convierte la tarea en un reto multidimensional:

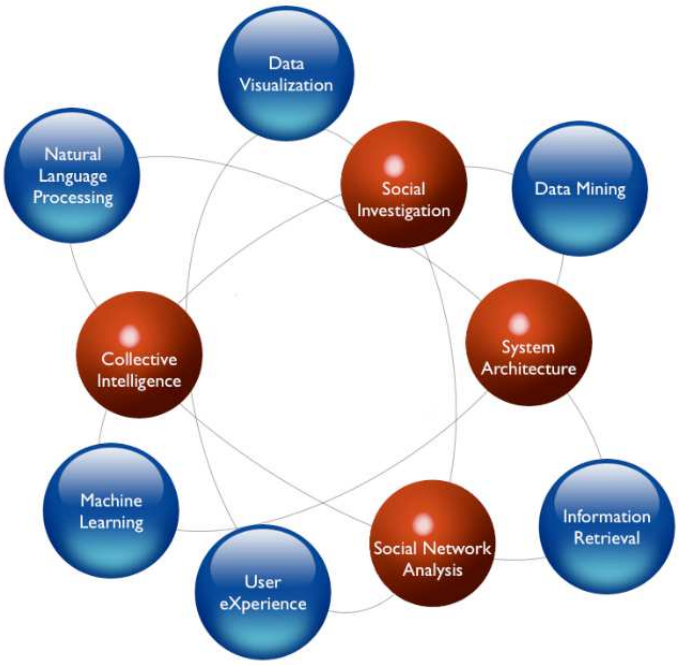

Figura 40. Reto multidimensional

Para resolver el reto con éxito se requiere una arquitectura tecnológica que proporcione la capacidad de trabajo con grandes cantidades de información, las denominadas Big Data, en tiempo real, integrando múltiples tecnologías y de manera totalmente escalable.

A continuación se presenta la arquitectura tecnológica de Cosmos para hacer frente a los retos anteriormente expuestos:

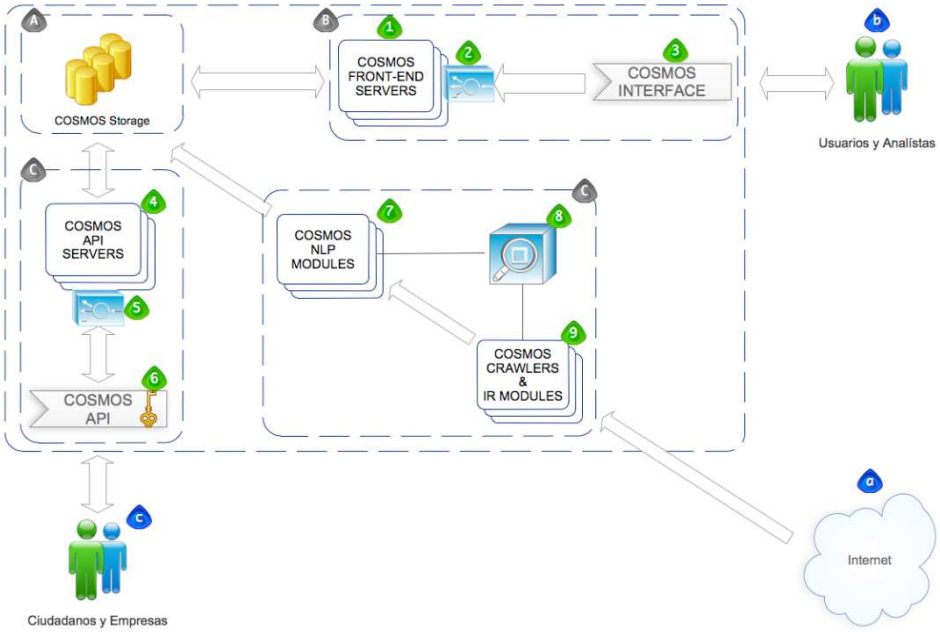

Figura 41. Arquitectura Tecnológica de Cosmos

Cosmos es una arquitectura modular, escalable, basada en *cloud computing*. En el anterior esquema, con letras mayúsculas se identifican los subsistemas que componen Cosmos y con números cada uno de los módulos de los subsistemas. Con letras minúsculas tenemos los diferentes agentes externos con los que Cosmos interactúa.

A continuación explicamos cada uno de estos subsistemas, los módulos que lo componen y las tecnologías utilizadas.

Como se podrá comprobar, la tecnología utilizada aún siendo puntera en la mayoría de los casos tiene un largo recorrido en el estado del arte de la investigación científica, así como una penetración importante en el mundo empresarial bajo proyectos como *Apache Fundation*. Es por ello que la tecnología utilizada, Open Source en su totalidad, es la mejor garantía de estabilidad de lo existente en el mercado.

- Agentes externos**:**

  - Internet: Es la fuente de información de la Escucha Activa y engloba los orígenes de información con los que trabaja Cosmos: medios de comunicación, blogosfera, foros, web, redes sociales, redes multimedia...

  - Usuarios y Analistas: Son los usuarios analistas de la Escucha Activa, tanto los analistas externos como los analistas de Autoritas Consulting, y dentro de esta categoría se engloban a los usuarios de la aplicación, los analistas de información, los *community manager*... Son por lo tanto usuarios

internos de la aplicación, que se validarán en la misma mediante usuario y contraseña, y que tendrán unos roles y unas capacidades de trabajo predefinidas.

- Ciudadanos y empresas: Son los usuarios externos de la aplicación que tendrán acceso a toda la información que el administrador haga pública a partir de la Escucha Activa. Accederán a un servicio que será registrado (*API Keys*) pero abierto a su uso con el objetivo principal de generar valor añadido al servicio.

- Subsistemas:

  - Subsistema de almacenamiento de datos: Encargado del almacenamiento de la información recuperada de internet y del proceso realizado sobre la misma. Los volúmenes de información son de tal magnitud que precisa de una integración de diferentes tecnologías que permitan su utilización en tiempo real y con una alta capacidad de respuesta.

  - Subsistema de Front End o aplicación Cosmos: Es el conjunto de módulos que proporcionan la interfaz de acceso a la aplicación y permiten interactuar con sus funcionalidades. Es un subsistema Web, SaaS, que permite el acceso a usuarios remotos con el simple uso de un navegador Web.

  - Subsistema API o de acceso externo: Cosmos permite la explotación de su base de escucha activa a partir de un acceso externo controlado vía API. Se permite acceder a todos sus datos raw y procesados de manera completa, permitiendo la construcción de aplicaciones de valor añadido sobre los datos. Todo el *front end* de la aplicación Cosmos podría ser reprogramada externamente con el uso de sus APIs. En definitiva, se provee de una infraestructura de Open Data que cumple el máximo nivel establecido:

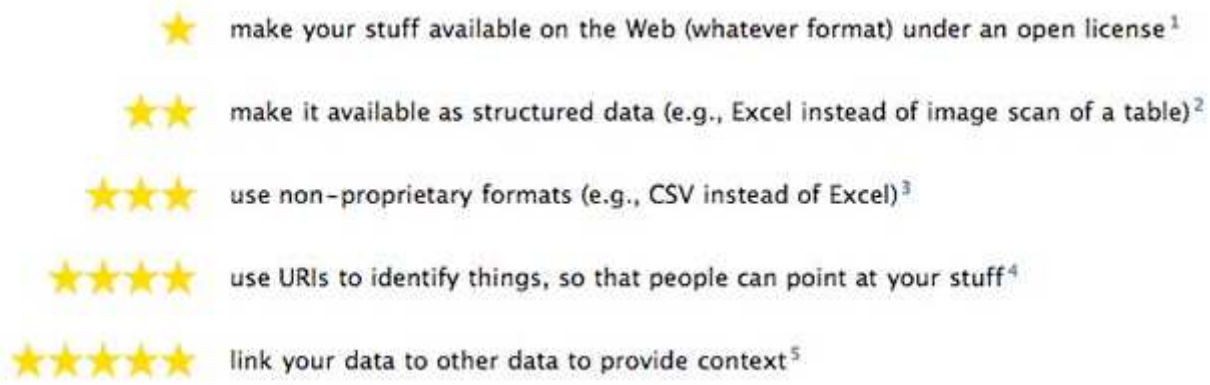

Figura 42. Niveles establecidos Open Data

- Subsistema de Proceso: Es el conjunto de módulos de procesamiento de Cosmos, que se dividen conceptualmente en dos grandes bloques, monitorizados para su autoescalado. Dentro de este subsistema, se encuentra el módulo descrito en este documento:

  - Cosmos Crawler & IR Modules: Módulos de recuperación de información desde Internet, encargados de transformar la definición estratégica del objeto de escucha en consultas específicas y especializadas en cada uno de los canales o fuentes de información, para su posterior recuperación y almacenamiento de la información.

Por lo tanto, para poder garantizar el tiempo real en el sistema de escucha activa, se debe garantizar en cada uno de sus módulos, para lo que se definen mediante un requisito de escalabilidad total.

Así pues, el sistema de recuperación de información, como se ha visto anteriormente, está descompuesto en diferentes módulos ligeramente relacionados, pero con la peculiaridad de que si uno de ellos no responde por algún motivo, el sistema en general puede seguir funcionando con total normalidad. Esto se ha realizado de esta manera debido principalmente a las limitaciones que presentan cada una de las APIs y pensando en la cantidad de proyectos que pueden llegar a estar ejecutados al mismo tiempo. Para conseguir esto, se ha utilizado un proceso basado en colas que ofrece al sistema una alta tasa de escalabilidad.

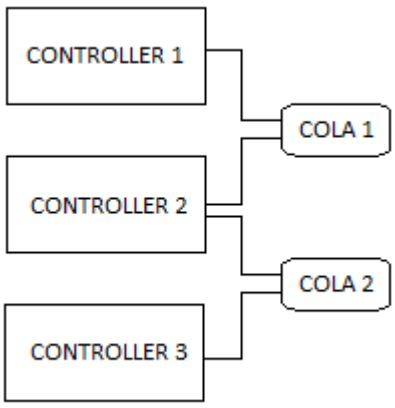

Figura 42. Sistema de colas

Dentro de las paralelizaciones utilizadas en el desarrollo se distinguen dos a nivel de aplicación: la de grano fino o la de grano grueso. Esta granularidad viene determinada por el tamaño en que se fracciona la aplicación, donde cada una de estas partes puede contener desde una única línea de código hasta una función o varias completamente definidas.

Así pues, la paralelización de grano fino se considera aquella en que los procesos son divididos en pequeñas porciones de código y los ciclos son fraccionados en subciclos que se ejecutarán de forma paralela, estableciéndose una comunicación bastante fluida entre ellos. Para lograr la máxima paralelización es necesario una arquitectura dedicada. Este tipo de granularidad esta enfocada a los datos.

Sin embargo, en la paralelización de grano grueso, las porciones de código divididas son mucho más grandes y complejas computacionalmente, reduciéndose considerablemente la comunicación entre ellos. El rendimiento de estos va ligado directamente con el número de procesadores o de las máquinas utilizadas. Como se puede apreciar, este tipo de granularidad esta enfocada a la paralelización de las tareas.

Por poner un ejemplo, en una paralelización fina de recuperación de noticias, tenemos una serie de computadoras procesando miles de páginas Web de forma paralelizada con una granularidad gruesa, mientras que dentro de cada una de estas máquinas tenemos a su vez una granularidad fina. A continuación, se muestra una tabla con los valores de los hilos ejecutados según el tiempo y los procesos.

THREADS	TIME	PROCESADOR/MEMORIA
1	902	10/50
5	285	15/55
10	155	20/55
50	48	100/60
100	49	100/65
500	55	100/70
1000	980	100/90

Figura 43. Tabla threads

Como se puede apreciar en la tabla, el número de hilos optimo en cuanto a tiempo de proceso es de 50, mientras que en la optimización del uso del procesador el óptimo es 10.

Además, para el procesado, escalado y paralelización de toda la información recuperada y almacenada se ha utilizado el *framework* MapReduce. Este sistema se compone de una función *map* y una *reduce*. La función *map*, se encarga de mapear los elementos de un modelo a otro. La función *reduce*, se encarga de disminuir una lista a un nuevo valor de otro modelo.

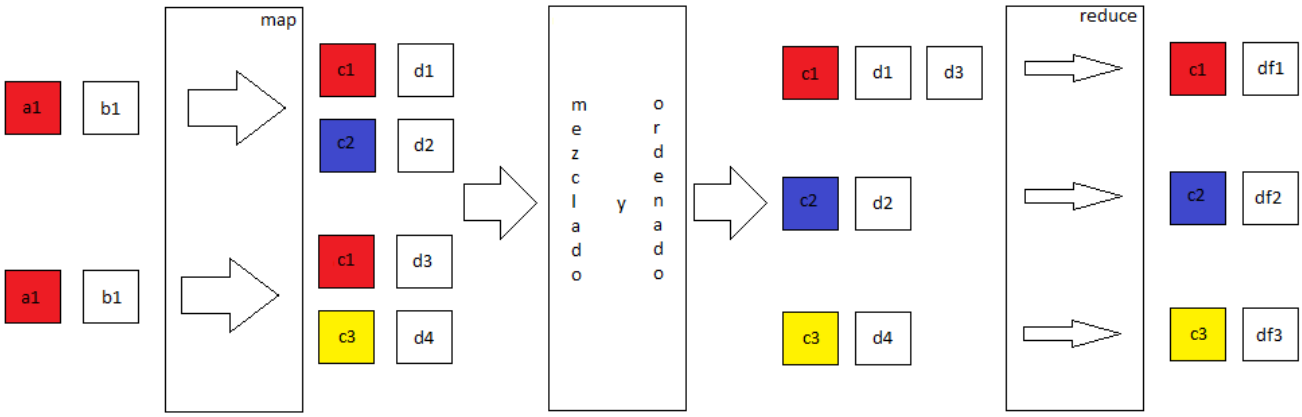

Figura 44. Esquema MapReduce

7. CONCLUSIONES

7.1 Soluciones del modelo

Internet se ha convertido en un medio de comunicación descontrolado y caótico, donde la información fluye de forma desestructurada de manera que para hacer una escucha activa se precisa de una herramienta de recuperación de información que obtenga la mayor cantidad posible y con una relevancia para el usuario final.

El sistema de recuperación de información presentado en el proyecto se ha desarrollado como modelo para mejorar la recuperación de información de Internet en la escucha activa. A diferencia de otros modelos, éste se ha basado en la utilización de aplicaciones de interfaces que facilitan la recuperación de información al sistema junto con la combinación de *crawlers* específicos que ayudan a enriquecer la información obtenida. De esta manera se puede aportar la mayor cantidad de información posible para un mejor análisis por parte del usuario, superando las expectativas que ofrecen otros sistemas de escucha activa en este ámbito.

Mientras que los sistemas de escucha activa presentados no distinguen entre diferentes tipos de Web, nuestro sistema diferencia una página según si es un blog o un foro, o en caso de no coincidir con alguno de los dos tipos, lo clasifica como una página simple. Una vez clasificado lo recupera según su naturaleza, facilitando la comprensión y el análisis al usuario. Además, divide la prensa entre lo que es prensa que existe también escrita y la que sólo es digital, lo que permite medir mejor el impacto en fases posteriores de la herramienta.

Por otra parte, la diversidad de los módulos implementados da más posibilidades de obtener mayores cantidades de información, además de presentar unos resultados más variados en función de la fuente seleccionada.

Además, otra ventaja de la modularidad propuesta es la facilidad de incorporación de nuevas fuentes de información. En la fecha de redacción de la tesina, se está trabajando en la incorporación de televisión, radio, podcast, Google+, Vimeo, Picasa o Pinterest.

7.2 Resultados y trabajo futuro

En el proyecto se ha implementado un sistema de recuperación de información de Internet que sirve como módulo indispensable para la escucha y que cuenta con algunas características que lo hacen especial.

- Integración de múltiples canales de información. Además de ser escalable y permitir la integración de nuevos canales en un futuro.
- Paralelización en muchas de las tareas, facilitando la ejecución de múltiples proyectos.
- Totalmente automatizado.
- Recuperación de mayor cantidad de información que la proporcionada por las interfaces, facilitando la comprensión de esta al usuario.
- Detección de qué tipo es la fuente de información.
- Tratamiento de contenidos en busca de la información más relevante y filtrada de código.
- Además de incorporar los nuevos canales de información mencionados anteriormente, se

puede crear listas de recuperación personalizadas por proyecto.

Se han ido elaborando varias líneas de trabajo en las que el proyecto puede ir evolucionando. Estas líneas se pueden ver a continuación:

- Con la aparición de las nuevas tecnologías y el éxito de las redes sociales y las aplicaciones de intercambio de contenidos, cada día surgen nuevas aplicaciones que pueden ser interesantes a la hora de recuperar información. Por lo que, la integración de esas nuevas interfaces puede darle un nuevo valor añadido al sistema de recuperación de información.

- Se plantea la incorporación de *crawling* mixto de información no estructurada (Nutch) con información estructurada (RSS).

- Implementación de *crawling* cruzado, es decir, la explotación de los enlaces que aparecen en un origen para recuperar información complementaria de otro origen, por ejemplo, los links a noticias que aparecen en Twitter, lo que además sirve como medida de influencia a módulos posteriores.

- Por último, otra línea de trabajo es la de la infraestructura paralela de recuperación consiguiendo una mayor y mejor integración de las bases de datos SQL y los sistemas de almacenamiento masivo no-SQL, así como el procesamiento paralelo basado en Mapreduce.

8. BIBLIOGRAFÍA

[1] Castells, M. Internet y la Sociedad Red. Conferencia de Presentación del Programa de Doctorado sobre la de Sociedad la Información y el Conocimiento. Universitat Oberta de Catalunya, 2000

[2] Arroyo Vázquez, Natalia. ¿Web 2.0? ¿Web social? ¿Qué es eso?. Educación y Biblioteca, núm. 161. http://eprints.rclis.org/archive/00011752/01/EYB_NA07.pdf, 2007

[3] Prosumidor. http://es.wikipedia.org/wiki/Prosumidor

[4] Paniagua Arís, Enrique. La gestión Tecnológica Del Conocimiento. ISBN: 978-84-8371-661-8. Universidad de Murcia, 2007

[5] Social Media. http://en.wikipedia.org/wiki/Social_media

[6] Theodore Holm, Nelson. http://es.wikipedia.org/wiki/Ted_Nelson

[7] Berners-Lee, Tim. http://es.wikipedia.org/wiki/Tim_Berners-Lee

[8] Cailliau, Robert. http://es.wikipedia.org/wiki/Robert_Cailliau

[9] Weblog. http://es.wikipedia.org/wiki/Blog

[10] Netscape Communications Corporation. http://en.wikipedia.org/wiki/Netscape

[11] Burbuja punto com. http://es.wikipedia.org/wiki/Burbuja_punto_com

[12] O'Reilly, Tim. What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software. Munich Personal RePEc Archive, MPRA Paper No. 4580, 2007

[13] Mermelada, Carlos A. Darwin y la teoría de la evolución. Disponible en: http://www.educarm.es/templates/portal/images/ficheros/etapasEducativas/secundaria/10/secciones/425/contenidos/13050/darwin_evolucion.pdf, 2009

[14] Wiki – http://es.wikipedia.org/wiki/Wiki

[15] Fundación Orange & TIGE. Estudio de uso del software social en la empresa Española. http://www.informeeespana.es/docs/softwaresocial_empresas2011.pdf, 2011

[16] Arthur, Charles. What is the 1% rule?. The Guardian (Technology section), 2006

[17] Trahtemberg, León. El impacto previsible de las nuevas tecnologías en la enseñanza y la organización escolar. Revista Iberoamericana de Educación, Número 24, 2000

[18] Folksonomía. http://es.wikipedia.org/wiki/Folcsonomía

[19] Blog. http://es.wikipedia.org/wiki/Blog

[20] Teoría de la información. http://es.wikipedia.org/wiki/Teoría_de_la_información

[21] Tecnologías de la información y la comunicación. http://es.wikipedia.org/wiki/Tecnologías_de_la_información_y_la_comunicación

[22] Toffer, Alvin. Future Schock. ISBN: 0-394-42586-3, 1970

[23] Lyman, Peter & Varian, Hal. Amount of new information doubled in last three years, UC Berkeley study finds, 2003

[24] Informació. http://es.wikipedia.org/wiki/Información

[25] Recuperación de información. http://es.wikipedia.org/wiki/Búsqueda_y_recuperación_de_información

[26] Salton, Gerald y McGill, Michel J. Introduction to modern information retrieval. ISBN: 0070544840. New York, 1983

[27] Croft, W. Bruce. Approaches to intelligent information retrieval. Information Processing and Management: an International Journal - Artificial Intelligence and Information Retriveal, Volume 23 Issue 4, Pages 249 – 254, 1987

[28] Meadow, Charles T. Text information retrieval systems. ISBN: 9780123694126. 1993

[29] Curso de verano de la UNED. Inteligencia Artificial en la Web, 2010

[30] Web 3.0. http://es.wikipedia.org/wiki/Web_3.0

[31] Sanagustín, Eva. Del 1.0 al 2.0: Claves para entender el nuevo marketing. ISBN: 978-84-9916-044-3, 2009

[32] Cosmos. http://cosmos.autoritas.net

[33] Robertson, K. Active Listening – More than just Paying Attention. Australian Family Physician, Vol. 34, (12) p. 994-1061, 2005

[34] Rangel Pardo, F.M.; Peñas, A. Clasificación de Páginas Web en Dominio Específico. Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN 89-96, 2008

[35] Rangel Pardo, F.M.; Peñas, A. Detecting Blogs Independently From the Language and Content, 1st. International Workshop on Mining Social Media, MSM09-CAEPIA09, 2009

[36] Rangel Pardo, F.M.; Buscaldi, D.; Rosso, P.; GIRPharma: A Geographic Information Retrieval Approach to Locate Pharmacies on Duty. 1st. International Conference on Computing for Geospatial Research & Application, COM.GEO, 2010

[37] Tromp, E.; Mykola, P. Graph-Based N-gram Language Identification on Short Texts. 20th Annual Belgian-Dutch Conference on Machine Learning, 2011

[38] Sala de prensa Online: Autorizada para todos los públicos. http://ivanpino.com/sala-de-prensa-on-line-autorizada-para-todos-los-publicos/, 2008

[39] Kaplan, R; Norton, D. The Balanced Scorecard: Meassures that Drive Performance. Harvard Busines Review, 1992

[40] Goleman, Daniel. Emotional Intelligence. ISBN: 0-553-37506-7, 1990

[41] Goleman, Daniel. Social Intelligence. ISBN-10:0-553-80352-2, 2007

[42] Mascardi, V.; Locoro, A.; Rosso, P. Automatic Ontology Matching via Upper Ontologies: A Systematic Evaluation. IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 5, pp. 609-623, 2010.

[43] Aguilar Gómez, A. Deformaciones de la Lengua Española en la Prensa. 1er. Congreso Internacional de la Lengua Española, 1997

[44] Esteban Rubio, A.; Menéndez Mañana, O. NER (Named Entity Recognizer). 3a. Conferencia de Procesamiento de Lenguaje Natural de la Universidad Europea de Madrid, 2003

[45] Armano, David. Six social media trends for 2012. Harvard Business Review. HBR Blog Network. 2012

[46] Liferay – http://www.liferay.com/documentation/liferay-portal/6.1/user-guide

[47] HTML – http://es.wikipedia.org/wiki/HTML

[48] CSS – http://es.wikipedia.org/wiki/CSS

[49] XML – http://es.wikipedia.org/wiki/Extensible_Markup_Language

[50] JSON – http://es.wikipedia.org/wiki/JSON

[51] Eclipse – http://es.wikipedia.org/wiki/Eclipse_(software)

[52] NetBeans – http://es.wikipedia.org/wiki/Netbeans

[53] Java – http://es.wikipedia.org/wiki/Java_(lenguaje_de_programaión)

[54] JavaScript – http://es.wikipedia.org/wiki/JavaScript

[55] JSP – http://es.wikipedia.org/wiki/JSP

[56] MySQL – http://es.wikipedia.org/wiki/MySQL

[57] Hibernate – http://es.wikipedia.org/wiki/Hibernate

[58] AIMC. Resumen general de resultados EGM. http://www.aimc.es/-Datos-EGM-Resumen-General-.html, 2012

[59] Rangel Pardo, Francisco M. & Anselmo Peña, Padilla. Detecting Blogs Independently from the Language and the Content. A. 1st International Workshop on Mining Social Media, MSM09-CAEPIA09, 2009

[60] Kohlschütter , Christian - Boilerplate Detection using Shallow Text Features - The Third ACM International Conference on Web Search and Data Mining New York City, NY USA, 2010

[61] Rangel Pardo, Francisco M. Análisis de Redes de Influencia en Twitter. CERI-12, 2012

9. ENLACES A APLICACIONES

[1] Wikipedia – http://es.wikipedia.org/wiki/Wikipedia:Portada

[2] del.icio.us – http://delicious.com

[3] Technorati – http://technorati.com

[4] Tripadvisor – http://www.tripadvisor.es/

[5] Digg – http://digg.com

[6] Aupatu – http://www.aupatu.com

[7] Wordpress - http://www.wordpress.com

[8] Blogger - http://www.blogger.com

[9] Blipback - http://www.blipback.com

[10] Twitter - http://www.twitter.com

[11] Google Reader - http://www.google.es/reader/

[12] Blogbridge - http://www.blogbridge.com

[13] GoogleMaps - http://maps.google.com

[14] Yahoo! Maps - http://espanol.maps.yahoo.com/

[15] Facebook - http://www.facebook.com

[16] Linkedin - http://www.linkedin.com

[17] Orkut - http://www.ourkut.com

[18] Xing - http://www.xing.com

[19] Flickr - http://www.flickr.com

[20] Youtube - http://www.youtube.com

[21] Slidershare - http://www.slidershare.net

[22] GoogleDocs - http://drive.google.com/start?authuser=0#home

[23] BrandCharts - http://www.brandchats.com/

[24] Radian6 - http://www.radian6.com/

[25] SocialMention - http://www.socialmention.com/

[26] Cosmos - http://cosmos.autoritas.net

[27] Liferay - http://www.liferay.com/

[28] CustomSearch API - https://developers.google.com/custom-search/v1/using_rest?hl=es

[29] Yahoo Search API - http://developer.yahoo.com/search/boss/boss_api_guide/webv2_service.html

[30] Bing Search API - https://datamarket.azure.com/dataset/5BA839F1-12CE-4CCE-BF57-A49D98D29A44

[31] YouTube API - https://developers.google.com/youtube/2.0/developers_guide_java

[32] Flickr API - http://www.flickr.com/services/api/

10. ANEXO

Como se ha mencionado anteriormente, el proyecto desarrollado es uno de los módulos indispensables para el producto de escucha activa de la empresa Autoritas Consulting.

Dicha empresa es una consultora especializada en estrategia de comunicación en internet, identidad digital y escucha activa. La compañía tiene una herramienta denominada *Cosmos* (anteriormente *Escucha Activa*) que es un mix de tecnologías de recuperación de información, procesamiento del lenguaje natural, análisis de redes sociales y aprendizaje automático para la generación de conocimiento estructurado a partir de la información desestructurada existente en las conversaciones producidas en internet.

Algunos ejemplos de repercusión mediática de los trabajos realizados son los siguientes:

- Noticia en el Diario digital de la agencia de noticias Telam sobre la monitorización con Escucha Activa de las Elecciones Argentinas 2011 (http://web.telam.com.ar/nota/5163/)

- Inclusión en la lista de 12 herramientas fundamentales de social media en "Una docena de..." (http://unadocenade.com/una-docena-de-herramientas-de-social-media-con-sabor-espanol/).

- Publicación del análisis de los efectos de las redes sociales en la política realizada con Escucha Activa en uno de los blogs sociopolíticos más influyentes de Chile (http://labs.utalca.cl/workshop/html/salaprensa.html y http://www.autoritas.es/2012/01/workshop-sobre-redes-sociales-y-elecciones-2012-en-la-universidad-de-talca-chile/)

- Publicación de resultados del imaginario colectivo de los candidatos políticos de las Elecciones Generales 2011 españolas en uno de los blogs sociopolíticos más representativos de habla hispana (http://www.cesarcalderon.es/?p=29542)

- Entrevista en el diario "El Centro" sobre el uso de herramientas como Escucha Activa para la realización de estrategias en el territorio político (http://www.cesarcalderon.es/?p=29692)

PÁG. 20

ENTREVISTA — EL CENTRO — Domingo 22 de enero de 2012

César Calderón, experto mundial en web 2.0, y "escucha activa":

"Los partidos políticos siguen funcionando como en los comienzos del siglo pasado"

Panelista del encuentro "Municipales 2012 y redes sociales", que se efectuará este miércoles 25 en Talca destacó que en las recientes elecciones argentinas "Twitter fue la plataforma más utilizada para la confrontación, muy por delante de los medios tradicionales o de Facebook"

25 de enero, a las 09.30 horas, comienza el seminario de social media en el auditorio de Medicina del Campus Lircay de la U. de Talca

¿Cómo fue su experiencia en la campaña de Cristina Fernández en Argentina? ¿Qué lugar ocupó Twitter? ¿Fue más central que otras redes sociales?

¿Cómo fue el estreno de la "escucha activa" de redes?

MUNICIPALES 2012

Algunos ven a Twitter como una burbuja, ¿qué piensa sobre esto?

La gran flexibilidad de los dispositivos conectados, como teléfonos y tabletas, supone más gente participando y en todo lugar...

¿Los medios tradicionales siguen teniendo más credibilidad...

Figura 45. Entrevista a César Calderón