UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

Departamento de Sistemas Informáticos y Computación

Universitat Politècnica de València

Análisis de datos colaborativos e inteligencia de negocio: aplicación al sector turístico

TESIS DE DOCTORADO

Tesis enviada para optar al grado de doctor de la Universitat Politècnica de València

NOVIEMBRE, 2020

Autor: Alexander Armando Bustamante Martínez

Directora: Eva Onaindía de la Rivaherrera

Directora: Laura Sebastiá Tarín

Dedicatoria

A mi abuela Fatme, por siempre creer en mi.

Agradecimientos

Son muchas las personas y entidades que aportaron, de diferente forma, para que este proceso de formación haya podido iniciar, continuar y terminar. Pero, quiero agradecer de manera especial a: mis tutoras, Eva y Laura, por su acompañamiento y paciencia en el desarrollo de esta tesis; mis papás, Alexander y Neris, por su apoyo en la decisión de emprender los estudios de doctorado; mis amigos, Paternina y Carbono, por su ayuda solucionando algunos inconvenientes técnicos; mi pareja, Camila, por comprender mis constantes ocupaciones; A Colciencias por el apoyo financiero que hizo posible el cursar el doctorado de manera presencial; y al Grupo de Tecnología Informática Inteligencia Artificial (GTI-IA) por acogerme durante el tiempo que estuve en la ciudad de Valencia.

Resumen

Desde hace varias décadas vivimos en lo que los académicos e industriales han convenido llamar la *era de la información y economía del conocimiento*, ambas caracterizadas, entre otras cosas, por el rol preponderante que ocupan tanto la información como el conocimiento en el quehacer y en los procesos, tanto productivos como de gestión, de las organizaciones. La información y el conocimiento han pasado de ser un recurso más en las organizaciones a ser uno de los principales activos que éstas poseen y utilizan para tomar decisiones, mejorar sus procesos, comprender el entorno y obtener una ventaja competitiva. Pero, para disfrutar de todos estos beneficios, se hace necesario una gestión pro-activa e inteligente de los datos.

Esta última se hace más necesaria en el contexto actual en donde la cantidad de datos disponibles sobrepasa la capacidad del hombre para analizarlos. Es en este contexto donde la Inteligencia de Negocios cobra especial importancia, ya que tiene como propósito tomar datos, generalmente, desde diferentes fuentes, integrarlos y procesarlos, dejándolos listos para posteriores tareas de análisis.

Paralelo al lugar importante que ocupa la inteligencia de negocios, está la contribución de la Web 2.0 en la generación de nuevo contenido. La Web 2.0 ha sido uno de los desencadenantes en la producción de datos a través de la internet convirtiéndose así en una fuente valiosa de datos sobre lo que las personas hacen, sienten y desean. Tal es el caso de plataformas como Twitter, que permite a las personas expresar su opinión sobre cualquier tema de interés u OpenStreetMap, que facilita la creación y consulta de información geográfica de manera colaborativa, entre otras.

Esta tesis gira en torno al uso de datos colaborativas y la utilización de la tecnología de la Inteligencia de Negocio para soportar el proceso de toma de decisiones, aplicado, concretamente, al sector turístico. Aunque el enfoque de tratamiento de los datos descrito en esta tesis puede ser utilizado, con ligeras adaptaciones,

para trabajar en otros dominios, se seleccionó el turismo por ser uno de las principales actividades económicas a nivel mundial. Tomando como referencia el año 2019, este sector económico creció en un 3.5 % por encima de la economía global que creció un 2.5 %, generó 330 millones de empleos (1 de cada 10) y representó el 10.3 % del producto interno bruto global.

En primer lugar, se realizó un análisis de las fuentes de datos colaborativas que pueden aportar conocimiento para el análisis de este sector y se seleccionaron cuatro fuentes de datos: OpenStreetMap y Twitter, ya nombradas y Tripadvisor y Airbnb para la información sobre alojamientos. Con las cuatro fuentes de datos colaborativas identificadas y utilizando la Inteligencia de Negocio como soporte tecnológico, se creó una plataforma responsable de todo el proceso, el cual abarca la extracción de datos de las diferentes fuentes, su integración en un formato consistente, su procesamiento y estructuración para ser utilizados en tareas de análisis y visualización de los resultados del análisis. La plataforma construida se denomina BITOUR.

BITOUR integra la propuesta de un modelo de BI para manejar datos geoespaciales, abiertos, combinados con contenido de redes sociales (colaborativos) junto con la propuesta de una serie de algoritmos para la identificación de los turistas y residentes de los destinos, la detección de usuarios no reales y la asignación de los tuits a los lugares dentro de un destino.

La integración de datos colaborativos, junto con los algoritmos, en una plataforma de Inteligencia de negocio representa una fuente potencial de valioso conocimiento que puede ser aprovechado en el sector turismo para conocer las actividades que realizan los turistas en un destino, las opiniones sobre un destino particular y sus atracciones, los periodos del año más frecuentados por los turistas según la nacionalidad, entre muchas otras preguntas.

BITOUR permite definir, interactivamente, un destino a analizar, cargar datos desde diferentes tipos de fuentes (espaciales y de opinión, entre otras), ejecutar rutinas que asocian opiniones a lugares e identifican turistas entre los datos reco-

pilados, así como visualizar los datos a través de la misma plataforma. BITOUR permite, entre otras cosas, la creación de tablas y gráficos dinámicos que posibilitan manipular los resultados de todos los cálculos que en la plataforma se han realizado. De esta manera, se pueden analizar tendencias de los turistas, tener un menor tiempo de respuesta frente a los eventos, enfocar mejor las campañas de mercadeo, etc. En definitiva, tener otra forma de acercarse a los turistas y comprenderlos.

Palabras clave: Almacén de Datos, Datos Abiertos, Fuentes Colaborativas, Inteligencia de Negocios, Turismo

Abstract

For several decades we have lived what academics and entrepreneurs call the information age and knowledge economy, both characterized, among other things, by the preponderant role that both information and knowledge hold in the production and management work of the organizations. Information and knowledge have evolved from being one among the resources in organizations to being one of their main assets in order to make decisions, to improve their processes, to understand the environment and to obtain a competitive advantage. But, to enjoy all these benefits, a pro-active and intelligent data management is necessary.

The latter is more necessary in the current context where the amount of available data exceeds human capacity to analyze it. It is in this context where Business Intelligence takes on special importance since its purpose is to take data, generally from different sources, integrate and process the data so as to leaving it ready for subsequent analysis tasks.

Parallel to the relevant role of Business Intelligence, there is the contribution of Web 2.0 in the generation of new data. Web 2.0 has been one of the triggers in the production of data through internet, thus becoming a valuable source of information about what people do, feel and wish. This is the case of platforms such as Twitter, which allows people to express their opinion on any topic of interest or OpenStreetMap, which facilitates the creation and consultation of geographic information in a collaborative way, among others.

This thesis revolves around the use of collaborative data and the use of Business Intelligence technology to support the decision-making process, specifically applied to the tourism sector. Although the data management approach described in this thesis can be used, with slight adaptations, to work in other domains, tourism was selected for being one of the main economic activities worldwide. Taking 2019 as a reference, this economic sector grew 3.5% above the global

economy, which grew 2.5%, generated 330 million jobs (1 in 10) and represented 10.3% of gross domestic product global.

First, an analysis of the collaborative data sources that can provide knowledge for the analysis of this sector was carried out and four data sources were selected: OpenStreetMap and Twitter, already mentioned, and Tripadvisor and Airbnb for information on accommodations. With these four collaborative data sources identified and using Business Intelligence as technological support, a platform responsible for the entire process was created, which includes the extraction of data from the different sources, integration of data in a consistent format, processing and structuring data to be used in analysis tasks and visualization of the analysis results. The built platform is called BITOUR.

BITOUR integrates the proposal of a BI model to handle open, geospatial data, combined with content from social networks (collaborative) together with the proposal of a series of algorithms for the identification of tourists and residents of the destinations, the detection of non-real users and the assignment of tweets to places within a destination.

The integration of collaborative data in a Business Intelligence platform represents a potential source of valuable knowledge that can be used in the tourism sector to know the activities that tourists carry out in a destination, the opinions about a particular destination and its tourist attractions or the seasons most frequented by tourists according to nationality, among many other questions.

BITOUR allows to interactively define a destination to be analyzed, to load data from different types of sources like spatial and opinion sources, to execute routines that associate opinions with places and to identify tourists among the collected data as well as visualize the data in the same platform. BITOUR allows for the creation of dynamic tables and graphics that make it possible to manipulate the results of all the calculations that have been performed on the platform. In this way, tourist trends can be analyzed to shorten response time to events,

put the focus on marketing campaigns, etc. In short, having another way of approaching tourists and understanding them.

Key words: Business Intelligence, Data Warehouse, Collaborative Data Sources, Open Data, Tourism

Resum

Des de fa diverses dècades vivim en el que els acadèmics i industrials han convingut dir la *era de la informació i economia del coneixement*, totes dues caracteritzades, entre altres coses, pel rol preponderant que ocupen tant la informació com el coneixement en el quefer i en els processos, tant productius com de gestió, de les organitzacions. La informació i el coneixement han passat de ser un recurs més en les organitzacions a ser un dels principals actius que aquestes posseeixen i utilitzen per a prendre decisions, millorar els seus processos, comprendre l'entorn i obtenir un avantatge competitiu. Però, per a gaudir de tots aquests beneficis, es fa necessari una gestió pro-activa i intel·ligent de les dades.

Aquesta última es fa més necessària en el context actual on la quantitat de dades disponibles sobrepassa la capacitat de l'home per a analitzar-los. És en aquest context on la Intel·ligència de Negocis cobra especial importància, ja que té com a propòsit prendre dades, generalment, des de diferents fonts, integrar-los i processar-los, deixant-los llestos per a posteriors tasques d'anàlisis.

Paral·lel al lloc important que ocupa la intel·ligència de negocis, està la contribució de la Web 2.0 en la generació de nou contingut. La Web 2.0 ha sigut un dels desencadenants en la producció de dades a través de la internet convertint-se així en una font valuosa d'informació sobre el que les persones fan, senten i desitgen. Tal és el cas de plataformes com Twitter, que permet a les persones expressar la seua opinió sobre qualsevol tema d'interès o OpenStreetMap, que facilita la creació i consulta d'informació geogràfica de manera col·laborativa, entre altres.

Aquesta tesi gira entorn de l'ús de dades col·laboratives i la utilització de la tecnologia de la Intel·ligència de Negoci per a suportar el procés de presa de decisions, aplicat, concretament, al sector turístic. Encara que l'enfocament de tractament de les dades descrit en aquesta tesi pot ser utilitzat, amb lleugeres adaptacions, per a treballar en altres dominis, es va seleccionar el turisme per ser un de les principals activitats econòmiques a nivell mundial. Prenent com a referència

l'any 2019, aquest sector econòmic va créixer en un 3.5 % per damunt de l'economia global que va créixer un 2.5 %, va generar 330 milions d'ocupacions (1 de cada 10) i va representar el 10.3 % del producte intern brut global.

En primer lloc, es va realitzar una anàlisi de les fonts de dades col·laboratives que poden aportar coneixement per a l'anàlisi d'aquest sector i es van seleccionar quatre fonts de dades: OpenStreetMap i Twitter, ja nomenades i Tripadvisor i Airbnb per a la informació sobre allotjaments. Amb les quatre fonts de dades col·laboratives identificades i utilitzant la Intel·ligència de Negoci com a suport tecnològic, es va crear una plataforma responsable de tot el procés, el qual abasta l'extracció de dades de les diferents fonts, la seua integració en un format consistent, el seu processament i estructuració per a ser utilitzats en tasques d'anàlisis i visualització dels resultats de l'anàlisi. La plataforma construïda es denomina BITOUR.

BITOUR integra la proposta d'un model de BI per a manejar dades geo-espacials, obertes, combinades amb contingut de xarxes socials (col·laboratius) juntament amb la proposta d'una sèrie d'algorismes per a: la identificació dels turistes i residents de les destinacions, la detecció d'usuaris no reals i l'assignació dels "tuits" als llocs dins d'una destinació.

La integració de dades col·laboratives en una plataforma d'Intel·ligència de negoci representa una font potencial de valuós coneixement que pot ser aprofitat en el sector turisme per a conèixer les activitats que realitzen els turistes en una destinació, les opinions sobre una destinació particular i les seues atraccions, els períodes de l'any més freqüentats pels turistes segons la nacionalitat, entre moltes altres preguntes.

BITOUR permet definir, interactivament, una destinació a analitzar, carregar dades des de diferents tipus de fonts (espacials i d'opinió, entre altres), executar rutines que associen opinions a llocs i identifiquen turistes entre les dades recopilades, així com visualitzar les dades a través de la mateixa plataforma. BITOUR permet, entre altres coses, la creació de taules i gràfics dinàmics que possibiliten

manipular els resultats de tots els càlculs que en la plataforma s'han realitzat. D'aquesta manera, es poden analitzar tendències dels turistes, tenir un menor temps de resposta enfront dels esdeveniments, enfocar millor les campanyes de mercadeig, etc. En definitiva, tenir una altra manera d'acostar-se als turistes i comprendre'ls.

Paraules clau: Dades Obertes, Fonts Col·laboratives, Intel·ligència de Negocis, Magatzem de Dades, Turisme

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1

Motivación

Este capítulo tiene como propósito describir los soportes en los cuales se sustentó esta tesis en términos de las necesidades que busca suplir y de las oportunidades que desea aprovechar (esto se presenta en la Sección 1.1). Luego, en la sección 1.2, se presentan cuáles son los objetivos que se alcanzaron con la culminación de esta tesis; en la Sección 1.3 se muestran qué aportaciones realiza la tesis tanto desde el punto de vista técnico como práctico. Para finalmente, en la sección 1.4, esbozar cómo está organizado el documento: qué capítulos tiene y el alcance de cada uno.

1.1 Justificación

Las Tecnologías de la Información y de la Comunicación (TICs) tienen omnipresencia en la vida de las personas y han conseguido hitos sin precedentes en su evolución. Hay tres hechos claves que han jugado un papel relevante en la evolución de las TIC en general y en el análisis de datos en particular:

- **Disponibilidad de datos.** Debido a la proliferación de dispositivos tecnológicos (particularmente teléfonos móviles), su intrusión en la vida de las personas y la evolución de la Web, hoy en día se tiene un ingente volumen de datos sobre casi cualquier aspecto, incluida la opinión, la percepción y el comportamiento de las personas. En un día se envían 500 millones de *tuits*,

294 mil millones de correos electrónicos y se crean 4 petabytes de datos en Facebook. Para el 2025, se estima que se crearán 463 exabytes de datos cada día en todo el mundo, lo que equivale a 212,765,957 DVD por día [122]. En la figura 1.1 se aprecia la cantidad de datos generados en algunas de las plataformas más populares en tan solo sesenta segundos. Este creciente volumen de datos esconde valiosa información y conocimiento.

- **Mejor capacidad de procesamiento y almacenamiento.** En consonancia con la conocida *Ley de Moore*[1], es posible tener progresivamente en menos tiempo unidades de procesamiento y almacenamiento más rápidas y con un menor costo. Esto ha llevado a disponer de dispositivos cada vez más potentes y más pequeños, además de posibilitar que más personas cuenten con dispositivos capaces de almacenar y procesar datos.

- **Democratización de las técnicas de análisis de datos.** Las tecnologías de análisis de datos, en especial el aprendizaje de máquina, han pasado de la fase de generación de expectativas a la fase de uso confiable; es decir, ha dejado de ser una promesa para convertirse en una realidad, permitiendo hacer usos predecibles a partir de su utilización. De este modo, han surgido herramientas que facilitan su uso, eliminando algunas barreras que impedían acceder a ellas y por consiguiente utilizarlas en ambientes productivos.

No cabe duda que nos encontramos en una época donde existe una gran cantidad de datos disponibles. Sobre este particular, hay dos aspectos destacables relacionados con los métodos para la creación y acceso a los datos, respectivamente. En cuanto a la creación de datos, la Web ha experimentado un cambio significativo donde los internautas han pasado de ser consumidores de información a creadores de la misma, fenómeno denominado Web 2.0, que permite a los usuarios en línea formar y participar en comunidades sociales para (co)-crear y distribuir contenido web. Un número creciente de usuarios de la Web participa

[1] La ley de Moore expresa que aproximadamente cada dos años se duplica el número de transistores en un microprocesador.

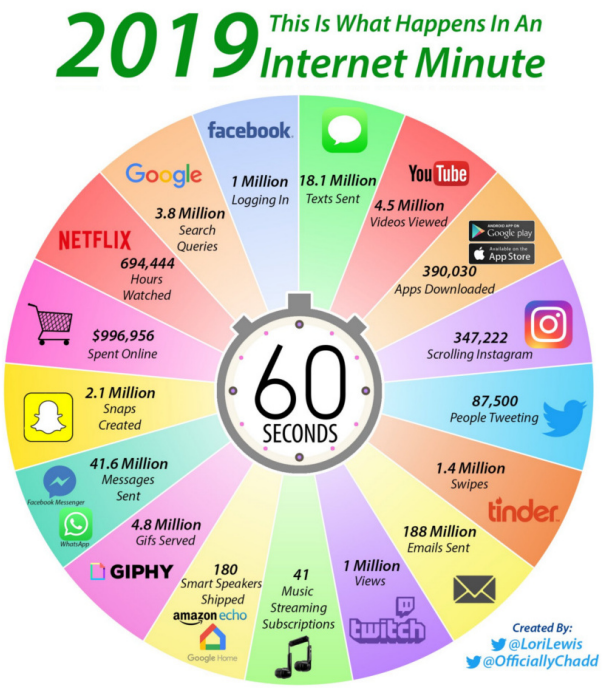


Figura 1.1: Qué ocurre en un día en Internet (figura tomada de [39])

en dicho intercambio de contenido y en actividades sociales en línea. En particular, los contenidos creados por los viajeros se perciben como altamente confiables, creíbles, relevantes, actualizados y atractivos. Los usuarios cada vez prestan más atención a los comentarios publicados en la Web, antes de tomar una decisión sobre, por ejemplo, una compra por Internet. Los usuarios afirman que se sienten seguros si comprueban los comentarios que se dejan en el sitio web antes de acudir a un hotel, restaurante, o atracción turística [160].

Por otro lado, bajo los principios de transparencia y compartición de datos que fomentan la innovación ha surgido un movimiento conocido como Datos Abiertos, cuyo objetivo es poner datos a disposición de la comunidad para ser utilizados gratuitamente. La creación colaborativa de datos y su acceso abierto representa una gran fuente de datos e información valiosa.

Paralelamente, *Business Intelligence* (BI), o Inteligencia de Negocios o Inteligencia Empresarial, como se denomina en castellano, es una tecnología que ha ganado aceptación durante la última década gracias a los beneficios que ofrece a las organizaciones para poder interactuar con los datos. La Inteligencia de Ne-

gocios permite tomar datos, tanto internos como externos, y transformarlos en información y conocimiento útil para el soporte a la toma decisiones [149].

Es en este contexto donde la BI gana relevancia como una forma de aprovechar los datos gratuitos y creados de manera colaborativa, disponibles en la Web con el fin de realizar análisis y ayudar en la toma de decisiones. Esta es la razón que ha motivado la construcción de una plataforma de BI, denominada BITOUR[2], cuyo objetivo es la recopilación de datos de diversas fuentes heterogéneas, el procesamientos de los mismos y su posterior depósito en un almacén de datos, dejándolos listos para realizar tareas de análisis.

Una solución de BI que utiliza fuentes de datos heterogéneas, abiertas y colaborativas constituye un atractivo esquema para crear ventaja competitiva de negocio en numerosos tipos de aplicaciones. Para este trabajo de tesis se ha seleccionado el dominio del turismo ya que este representa una de las principales actividades económicas a nivel global. Tomando como referencia el año 2019, el sector turístico creció en un 3.5 % por encima de la economía global que creció un 2.5 %, generó 330 millones de empleos (1 de cada 10) y representó el 10.3 % del producto interno bruto global.

La selección del sector turístico es especialmente relevante para este trabajo porque esta tesis de investigación tiene su origen en la mejora de la competitividad del sector turístico de la ciudad de Santa Marta, Colombia, país clasificado como "en desarrollo" según las Naciones Unidas [107]. El turismo en Santa Marta, Colombia, es la principal actividad económica y es el tercer destino elegido por los turistas nacionales en temporadas de vacaciones, como Semana Santa. El interés por incentivar el turismo de esta región ha llevado a sus dirigentes a la creación de un programa de formación específico en este ámbito, con inclusión de becas de doctorado, del que el autor de esta tesis es unos de los beneficiarios.

La industria turística se enfrenta a un reto, llegar al consumidor aprovechando los recursos que ofrece internet y buscando espacios de interacción con sus

[2]Disponible: `http://demo.softsimulation.com/`

consumidores para mostrarles sus productos y estrechar relaciones. El papel de las redes sociales es fundamental. Esto ha dado lugar al fenómeno del viajero 2.0 [92], aquel que recomienda, comparte en tiempo real fotografías, sentimientos e información y experiencias relevantes –tanto positivas como negativas- de los servicios turísticos. El uso de las redes sociales está siendo determinante en una nueva relación entre proveedores y clientes, está innovando al sector la forma de comunicación y la promoción turística de servicios. La libre circulación de opiniones en las redes sociales, es uno de los fenómenos que mejor definen el nuevo marketing interactivo [161]. La cantidad de fuentes disponibles es amplia y las hay de dominio general como Twitter y Wikipedia, de dominio específico como Tripadvisor y de orden internacional como es el caso del reporte de la competitividad turística publicado por el Foro Económico Mundial [48].

Por consiguiente, se cuenta con una variedad de fuentes de datos que pueden ser aprovechadas por BITOUR. Cada una de ellas con datos relevantes que pueden ser explotados en términos de entender mejor el sector turístico y tener información que soporte la toma de decisiones.

Desde una perspectiva computacional, esta tesis centra sus aportes en dos puntos: 1) la integración de datos: tanto desde diferentes fuentes como de diferente tipo (espaciales y colaborativos). 2) Hace uso de estos datos integrados para presentar propuestas de algoritmos para la asignación de los tuit a lugares dentro del destino, para la identificación de usuarios no reales y para la clasificación entre turistas y residentes.

Si bien los dos primeros algoritmos son importantes para el funcionamiento de la herramienta, el último, centrado en la identificación de los turistas, representa un aspecto clave para todas las tareas de análisis que permite hacer la herramienta. Por esta razón, se centró gran parte del esfuerzo en ir un poco más allá de los enfoques tradicionales y se incluyó un algoritmo de aprendizaje de máquina que arrojó importantes resultados como se describirá en la sección 6.1.3.

Esta tesis en síntesis y como muestra la figura 1.2 permite tomar datos colaborativos y abiertos de utilidad para el sector turístico, insertarlos en un almacén de datos centralizado que permite su posterior procesamiento y análisis en función de los usuarios catalogados como turistas y sus tuits, buscando dar respuestas a preguntas como: ¿quiénes son turistas? ¿cuál es el tiempo total de su estadía?, ¿cuántas atracciones visitaron?, ¿qué opinión tiene sobre las atracciones?, entre otras.

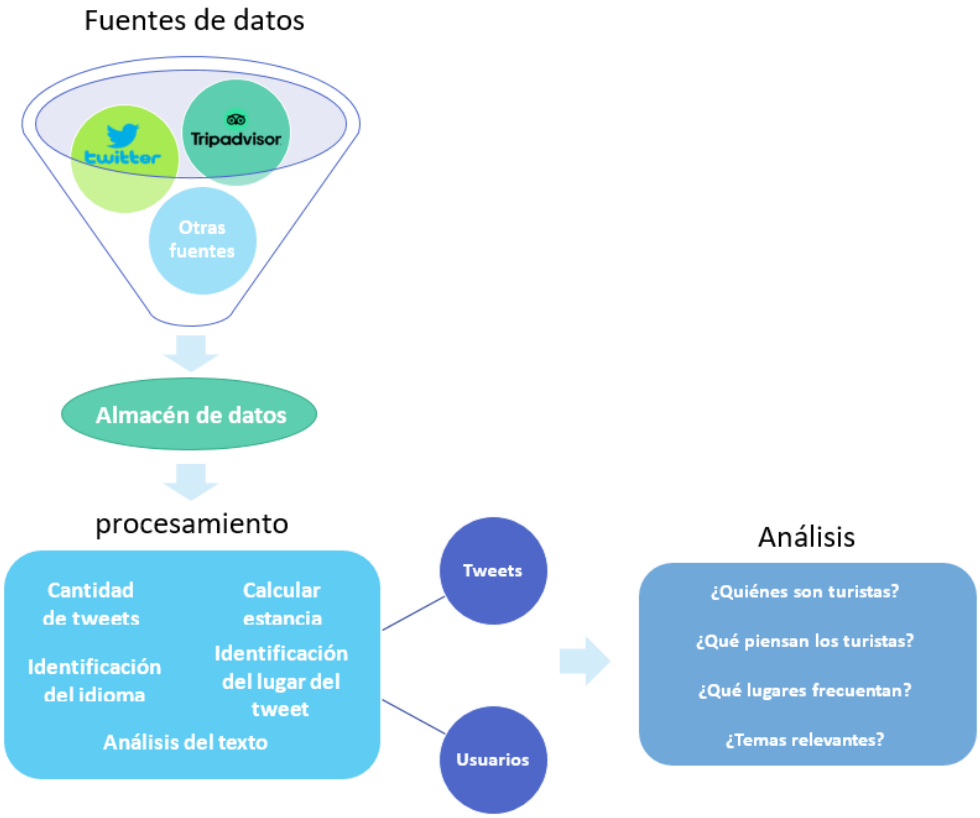


Figura 1.2: Resumen gráfico de la propuesta

Todo lo descrito, para cumplir el fin último de BITOUR: soportar el proceso de toma de decisiones en el ámbito del turismo, a través de la posibilidad de: evaluar tendencias con respecto a las preferencias de los turistas; permitir tener un mayor tiempo de respuesta frente a sucesos que pueden pasar ocultos como lo son: ¿qué están pensando los turistas del destino?, posibilitar el marketing segmentado al conocer las diferencias que existen entre algunos tipos de turistas, etc.

1.2 Objetivos

El objetivo general de esta tesis es desarrollar una solución de Inteligencia de Negocios que incorpore información colaborativa y abierta para soportar la toma de decisiones, tomando como escenario de ejemplificación su aplicación en el sector turístico.

Los objetivos particulares que se persiguen son:

- **Identificar y categorizar las fuentes de información relevantes.** Existe variedad de fuentes disponibles en la Web tanto de carácter colaborativo como abierto, por lo que inicialmente se procede a identificar de la plétora de fuentes disponibles, cuáles pueden ser de interés para la tesis. Se identifica el tipo de datos que proporciona cada una de ellas, el formato, la frecuencia y el proceso de extracción de los datos.

- **Idear la forma en que los datos desde las diferentes fuentes pueden ser integrados de manera consistente para analizar el dominio del turismo.** Manipular datos desde diferentes fuentes impone el reto de: cómo integrar datos desde fuentes dispares y mantener la consistencia semántica.

- **Comprobar si es posible utilizar datos de fuentes colaborativas y públicas para estructurar una solución de Inteligencia de Negocios.** Una vez las fuentes relevantes son identificadas es necesario verificar si es posible articularlas con una solución de Inteligencia de Negocios que permita el análisis del sector turístico.

- **Proponer un enfoque para la asignación de tuits a los lugares y para la identificación de los turistas.** Dos aspectos claves para el análisis del turismo son: saber quiénes son los turistas y qué lugares visitan. Por lo que se debe buscar una forma práctica que permita, a partir de los datos colaborativos y públicos, saber quienes son turistas y qué lugares visitan dentro de un destino.

- **Desarrollar la plataforma de visualización.** Integrar las funcionalidades construidas en una aplicación con un componente de visualización que permita una vez calculada la información y conocimiento de interés, desplegar y ponerlo a disposición de quienes puedan estar interesados.

1.3 Aportaciones

Las principales aportaciones del presente trabajo se resumen a continuación:

1. Estado del arte con respecto al uso de la Inteligencia de Negocios en diversos dominios de aplicación, haciendo énfasis en el turístico.

2. Análisis de las fuentes de datos colaborativas y abiertas que pueden proporcionar datos relevantes para soportar la toma de decisiones que contribuyan a mejorar la competitividad turística de los destinos.

3. Una plataforma Web flexible que permite analizar la afluencia de los turistas a un destino así como su percepción del mismo y los sitios visitados dentro de este.

4. Implementación de rutinas que permiten a partir de las fuentes colaborativas y abiertas utilizadas la asignación de los usuarios a lugar dentro del destino y la posterior identificación de los turistas.

1.4 Organización del documento

- **Capítulo 1. Motivación**

  Es este capítulo se presentan los objetivos y la motivación del trabajo de tesis, junto con sus principales aportaciones. Por último, se incluye el presente apartado en el que se muestra la organización de esta memoria.

- **Capítulo 2. Inteligencia de Negocios**

  Es este capítulo se presentan los sustentos conceptuales que soportan este

trabajo de tesis, girando alrededor de lo que es la Inteligencia de Negocios, cuáles son sus componentes y qué aplicación tiene.

- **Capítulo 3. Metodología**

  En este capítulo se describe cuál fue el proceso seguido para el cumplimiento de los objetivos de esta tesis. Es decir, se aborda cada una de las fases y las decisiones que fueron tomadas en cada una de ellas.

- **Capítulo 4. Diseño de BITOUR**

  En este capítulo se presenta cómo están organizados cada uno de los componentes que conforman la solución de Inteligencia de Negocios construida. Es decir, cómo estos componentes hacen posible las prestaciones de las funcionalidades requeridas.

- **Capítulo 5. Fuentes de datos**

  En este capítulo se describen aquellas fuentes de datos que sirven de proveedores de datos para este trabajo. Se describe precisamente cuáles son las fuentes, qué datos proveen y cómo se puede acceder a estos de manera gratuita.

- **Capítulo 6. Procesamiento y visualización de datos**

  En este capítulo se describe por un lado, aquellas rutinas que demandan un mayor nivel de detalle debido a que entender cómo funcionan es indispensable para entender la plataforma construida; y por otro, las funcionalidades de las herramienta construida relacionadas con la visualización de datos, abarcando qué se puede realizar con la herramienta y cómo.

- **Capítulo 7. Conclusiones y trabajo futuro**

  En este capítulo se exponen las aportaciones más relevantes de este trabajo de tesis y se indican algunas de las futuras líneas de investigación por las que puede continuarse este trabajo.

CAPÍTULO 2

Inteligencia de Negocios

En este capítulo se describe uno de los conceptos principales en torno al cual gira la presente tesis: la Inteligencia de Negocios. Para esto se inicia con la sección 2.1 presentando diferentes acepciones que se han hecho del término y cómo este ha evolucionado a través del tiempo; luego en la sección 2.2 se ilustra la arquitectura típica sobre la cual se suele organizar un sistema de Inteligencia de Negocios y los diferentes componentes que hacen parte de cada una de las capas de esta arquitectura; después, en las secciones 2.3 y 2.4 se presentan las aplicaciones que se han hecho utilizando la Inteligencia de Negocios en diferentes contextos como son: el sector bancario, el sector de la salud y el sector del turismo; finalmente, la sección 2.5 se presentan trabajos que también buscan realizar la integración desde diferentes fuentes.

2.1 ¿Qué es la Inteligencia de Negocio?

Una de las definiciones más recurrentes y destacadas del concepto de Inteligencia de Negocios (BI por sus siglas en inglés) es la propuesta por Howard Dresner en 1989 cuando acuñó el término para referirse a un concepto sombrilla que engloba varios elementos y métodos enfocados a mejorar el proceso de toma de decisiones, utilizando sistemas basados en hechos [164]. Los sistemas de Inteligencia de Negocios proporcionan información procesada que se entrega en el momento adecuado, en la ubicación correcta y en el formato correcto para ayudar

en la toma de decisiones [119]. Existen múltiples definiciones de BI, algunas de las cuales se muestran en la Tabla 2.1, que conjuntamente permiten tener una visión más amplia sobre las diferentes acepciones del término. Por ejemplo, en la Tabla 2.1 se aprecia que la consultora Gartner, una de las empresas líder mundial en consultoría tecnológica, incorpora en la definición de BI no solo la tecnología sino el conjunto de buenas prácticas que permiten cumplir con el propósito de "brindar acceso al análisis de la información para mejorar y optimizar las decisiones y el rendimiento".

Autor	**Definición**
TDWI	Es un espectro completo de tecnologías que consisten en múltiples productos complementarios, que van desde consultas simples hasta análisis OLAP y análisis predictivo como la minería de datos aplicada [41].
Forrester	Un conjunto de metodologías, procesos, arquitecturas y tecnologías que transforman los datos sin procesar en información significativa y útil que se utiliza para permitir una percepción de la organización y tomar decisiones estratégicas, tácticas y operativas más efectivas [47].
Gartner	Es un término general que incluye las aplicaciones, la infraestructura, las herramientas y las mejores prácticas que permiten tener acceso al análisis de la información para mejorar y optimizar las decisiones y el rendimiento [55].
Negash y Gray	Los sistemas de Inteligencia de Negocios combinan la obtención y almacenamiento de datos, así como la gestión del conocimiento con herramientas analíticas que presentan información compleja y competitiva a los planificadores y tomadores de decisiones [108].

Tabla 2.1: Definiciones de BI

Aunque la definición original de BI de Dresner, así como las definiciones más recientes de analistas como Gartner, Forrester y TDWI [41, 47, 55, 108] son de amplio alcance, también suele utilizarse en un sentido más estrecho, asociado a un conjunto limitado de capacidades que incluye (a) el proceso de extracción, transformación y carga de los datos; (b) el almacenamiento de datos en repositorios centralizados; (c) el procesamiento analítico de datos en línea; (d) y, en algunos casos, técnicas de análisis predictivo y generación de informes [74, 70].

En determinados contextos el término BI se utiliza indistintamente para hacer referencia a otros términos relacionados como la gestión de grandes volúmenes

de datos (*Big Data*) o minería de datos. Específicamente, un sistema de Inteligencia de Negocios está guiado por el propósito de ofrecer soporte para la toma de decisiones y, para cumplir con este propósito, utiliza un amplio número de herramientas tales como consultas de bases de datos históricas, externas e internas, así como técnicas de *Big Data* y minería de datos.

Se pueden distinguir dos enfoques principales en la Inteligencia de Negocios, uno de los cuales hace hincapié en la parte gerencial y/o administrativa, y otro en la parte tecnológica. En el siguiente apartado se describirán brevemente estos dos enfoques.

2.1.1. Enfoques a la Inteligencia de Negocios

Al hablar de BI se suelen utilizar dos prismas o perspectivas diferentes, aunque no opuestas. Por una lado está la perspectiva administrativa o gerencial, la cual se sustenta en la estrategia empresarial para su articulación. Por otro lado, el enfoque tecnológico, basado en las herramientas y/o tecnología que dan soporte al proceso de BI. De este modo, las competencias requeridas pueden subdividirse en competencias de gestión y competencias técnicas (metodológicas, conceptuales y específicas de los productos utilizados) [52]. En la Tabla 2.2 se ofrece una breve descripción de las características de estos dos enfoques y se amplía la descripción en las secciones siguientes.

Enfoque	Características
Administrativo	Centrado en el proceso, concibiéndolo organizado y sistemático. Por el se adquieren, analizan y disemina información de fuentes de información internas y externas significativas para sus actividades comerciales y para la toma de decisiones. Permite revelar hallazgos sobre la organización en sí misma y su relación con sus mercados, clientes, competidores y economía.
Técnico	Centrado en las herramientas. Presenta BI como un conjunto de herramientas que respaldan el proceso descrito anteriormente. El enfoque no está en el proceso en sí, sino en las tecnologías que permiten la recuperación, manipulación y análisis de la información

Tabla 2.2: Enfoques BI

2.1.1.1. Enfoque administrativo

El enfoque administrativo interpreta la Inteligencia de Negocios como un proceso que busca alinearse con la estrategia empresarial y traducirla en elementos que permitan tomar decisiones rápidas y acertadas. El objetivo de este enfoque es dotar a las empresas de las capacidades necesarias para llevar a cabo un proceso de gestión basado en información que permita describir el ambiente competitivo, pronosticar el contexto futuro, desafiar las suposiciones subyacentes haciendo las preguntas correctas, identificar y compensar las debilidades expuestas, utilizar la inteligencia para implementar y ajustar la estrategia a los cambios del entorno y determinar cuándo la estrategia ya no es sostenible [66].

Para dar cumplimiento a los objetivos del enfoque administrativo se requiere la selección y aplicación de marcos, modelos, métodos y metodologías a partir de las cuales se produce una decisión en un nivel organizativo particular, variando la información que se presenta a cada persona en función de sus necesidades. Algunas de las herramientas utilizadas en este enfoque son [66]:

- Escaneo del entorno enfocado en el reconocimiento de factores externos e internos de las organizaciones, priorizando la actividad de mercado sobre las demás funciones organizativas.

- Inteligencia competitiva enfocada principalmente en la identificación y análisis de los competidores.

- Estrategia competitiva para priorizar el poder de negociación con los compradores o clientes, con los proveedores o vendedores así como la amenaza de nuevos competidores y de productos sustitutos, y la rivalidad entre los competidores.

- Metodologías de gestión como el *balanced scorecard* para construir un sistema de indicadores basados en la información de la organización.

- Técnicas de gestión del rendimiento, costes basados en actividad y gestión por procesos de negocio.

- Paradigmas de comprensión organizativa como el pensamiento sistémico y el pensamiento complejo.

2.1.1.2. Enfoque tecnológico

En su dimensión tecnológica, el objetivo de la Inteligencia de Negocios es la integración de los datos generados por y para la organización, procesar los datos de manera que sirvan como entrada a diferentes procesos de gestión en cada uno de los niveles organizativos, y distribuir la información generada para los usuarios interesados.

Desde esta perspectiva, es necesario integrar varios elementos para obtener una dimensión práctica de la BI. Entre estos elementos se encuentran las bases de datos de procesamiento transaccional, bases de datos analíticas, la minería de datos, los sistemas de generación de informes y la visualización de datos. Cada uno de estos elementos se desarrolla independientemente para dar solución a problemáticas específicas en diferentes momentos temporales y por ende su uso se puede aplicar de forma aislada. No obstante, como se infiere del objetivo fundamental de la Inteligencia de Negocios, su utilización de forma integrada permite constituir el sistema de BI aún cuando cada una de estas herramientas no constituyen en sí mismas "una solución" de BI.

Existen varios trabajos que presentan una tipología descriptiva y categorizada de las habilidades requeridas para un profesional en BI [52]. Las habilidades se agrupan en siete categorías que son: (1) preparación de los datos para expertos en la materia; (2) aplicación de técnicas de modelado y simulación así como de técnicas estadísticas para descubrir nueva información; (3) gestión de las partes interesadas; (4) desarrollo de una hoja de ruta estratégica de BI a largo plazo donde se vincule la estrategia de la empresa; (5) comprensión de los procesos

de negocio con el fin de extraer los requisitos del usuario de forma efectiva; (6) diseño de soluciones sostenibles; (7) extracción y distribución de conocimiento.

Finalmente, es importante destacar algunas variantes que una solución de BI puede adoptar con respecto a su implementación [149]:

- **Operacional**. La presión competitiva de las empresas actuales ha aumentado la necesidad de un BI casi en tiempo real (también denominado BI operativo). El objetivo del BI operativo es reducir la latencia entre el momento en que se adquieren los datos de la transacción y el momento en que esos datos están disponibles para su análisis, de modo que se puedan tomar las medidas adecuadas cuando se produce un evento. Con este objetivo, las empresas desean detectar patrones o las tendencias temporales sobre la transmisión de datos operativos de los eventos.

- **Situacional**. El aspecto situacional hace referencia a la toma de conciencia por parte de las empresas de los eventos que ocurren en el mundo y que pueden afectar sus negocios (por ejemplo, comentarios positivos o negativos sobre sus nuevos productos, desastres naturales, etc.). Habitualmente esta información externa no está estructurada y debe integrarse con información interna del almacén de datos.

- **Autoservicio**. Este tipo de análisis de datos permite a los usuarios finales crear consultas analíticas e informes sin la necesidad de la participación del departamento TIC. La interfaz de usuario en aplicaciones de autoservicio debe ser intuitiva y fácil de usar de modo que las personas que no tengan conocimientos técnicos puedan acceder y trabajar con la información corporativa. Adicionalmente, el usuario también debe tener la posibilidad de incluir más fuentes de datos y complementar así los datos ya disponibles.

- **En memoria**. Este nuevo enfoque de BI se centra en cargar todo o gran parte de las estructuras de datos que soportan los sistemas de inteligencia de negocios en la memoria principal del computador, en oposición, o comple-

mentariamente, al enfoque tradicional, centrado en cargar la mayor parte de los datos en disco. Proporciona un acceso más rápido a la información que los sistemas almacenadas en disco ya que los algoritmos de optimización internos son más simples y usan menos instrucciones de CPU, proporcionando de este modo un mayor rendimiento de las consultas. No obstante, este enfoque tiene limitaciones relacionadas con la cantidad de datos que puede almacenarse, la persistencia y el coste.

2.1.2. Evolución del concepto BI

Existe acuerdo en la literatura académica que la primera mención al concepto Inteligencia de Negocios fue en el año 1958 y la realizó el investigador de la IBM Hans Peter Luhm [94]. Luhm define BI como ”la capacidad de comprender las interrelaciones de los hechos presentados de tal manera que guíe la acción hacia una meta deseada”. Adicionalmente, en el citado trabajo se especifica un sistema de BI como un sistema automático desarrollado para difundir información a diversas secciones de cualquier organización, científica o gubernamental. En el enfoque de Luhn, el sistema se centra en la extracción automática de documentos y en la entrega de esta información a los puntos de acción apropiados. En otras palabras, un sistema que disemina información automáticamente, clasificándola y enviándola a diferentes lugares, según criterios definidos por los usuarios.

Otro pionero en la creación del término BI, fue Richard Greene, quien en el año 1966 lo definió como "la información procesada de interés para la administración acerca del presente y futuro del entorno en el cual el negocio debe operar"[59]. Esta aproximación aborda el concepto BI como una forma de espionaje, incorporando conceptos de inteligencia militar.

Si bien las primeras menciones al termino de Inteligencia de Negocios las realizaron los autores Luhn y Greene en los años 60, es la definición de Howard Dresner realizada 30 años después en el año 1989 la que mayor divulgación y aceptación ha tenido en la comunidad [119]. Esta definición, como se ha comen-

tado en la sección 2.1, considera a la BI como un termino sombrilla que engloba a un conjunto de técnicas que van desde tecnologías y metodologías hasta buenas prácticas para la toma de decisiones.

Como se muestra en la Tabla 2.3 el concepto BI ha ido evolucionando con la aparición de tecnologías de almacenamiento, análisis y procesamiento de datos [25]. En la década de los 80 y principio de los 90 la Inteligencia de Negocios se centró en recolectar información, posiblemente descentralizada, con el objetivo de analizarla de forma manual. En la década de los 90 la información se centraliza y se complementa con el uso de procesamiento analítico en línea, herramientas de informes y otras herramientas de análisis como minería de datos, a veces denominadas herramientas de analítica de datos. Finalmente, en la primera década del nuevo siglo se aumentó la capacidad de almacenamiento y procesamiento de la información y también surgió la posibilidad de almacenar el conocimiento tanto implícito como explicito que mueven las organizaciones.

En los últimos años, los sistemas de BI han experimentado un cambio: las redes sociales, sensores de máquinas, dispositivos como teléfonos inteligentes y otras fuentes generan nuevos datos que a menudo difieren de los datos operativos tradicionales con respecto a su estructura, tasa de crecimiento y volumen [101, 116]. Es así como el almacenamiento, procesamiento y análisis de grandes cantidades de datos ha dado lugar al área *Big Data*. Se utiliza el concepto *Big Data* para referirse a grandes conjuntos de datos que no caben en una sola memoria. Pero es un error asociar el término a un volumen fijo de datos ya que *Big Data* no es una noción estática; es decir, lo que hoy se considera *Big Data* quizás no lo será en pocos años como consecuencia de que habrá mejores capacidades de almacenamiento y de procesamiento. En pocas palabras, *Big Data* es un refuerzo de BI.

Época	Hecho relevante	Descripción
Inicio de los 80s	Delimitación de las fuentes de información	Caracterizada por la amplia influencia del modelo de las cinco fuerzas de Porter. Definió que para la organización es relevante contar con datos de cliente, productos, competidores y proveedores.
Mediados de los 80s	Necesidad de predecir comportamientos futuros	Inclusión de las teorías de escenarios de pronósticos que permiten hacer un estudio sistemático de futuros posibles, probables y preferibles
Finales de los 80s	Incorporación de la Psicología como herramienta de análisis	Se incorporaron elementos psicológicos propuestos por Myers-Briggs. Se orientara el análisis al del perfil del competidor.
Inicio de los 90s	Dominio de los sistemas de información gerenciales y transición de los EIS a los ERP	Se hizo evidente otra orientación que dio gran prioridad a la información disponible para la alta dirección. A comienzos de los 90s se cambió de los Sistemas de Información Empresariales hacia los Sistemas de Planeación de Recursos Empresariales que permitieron tener más información de los procesos de la empresa.
Mediado de los 90s	Incorporación de los almacenes de datos	BI se asoció con las herramientas tecnológicas que permiten recopilar y visualizar dinámicamente los datos contenidos en los almacenes de datos.
Finales de los 90s	Incorporación de las minería de datos	Para finales de los años 90 se incluyó una poderosa herramienta analítica como parte de BI: la minería de datos, que permitió usar los datos para extraer conocimiento relevante.
2000s	Incorporación de la Gestión del conocimiento	Recordó a los analistas que las tecnologías no son suficientes para conseguir la inteligencia de negocios: es necesario que la unan a su propia experiencia.
2013	Incorporación del Big Data	Ante el creciente volumen de datos y la diversidad en los mismos, hubo la necesidad del uso de las nuevas tecnologías orientadas al almacenamiento y procesamiento de estos.

Tabla 2.3: Evolución de la BI

2.1.3. Beneficios de las soluciones BI

La creciente necesidad por el aprovechamiento de los datos y la información organizativa unido a la creciente disponibilidad de nuevos datos dejan un terreno abonado para alternativas como la Inteligencia de Negocios que se fundamentan en el análisis de datos [77, 28]. Una función de las herramientas BI es recopilar hechos en bruto no organizados (datos) y ponerlos en un contexto específico (información). La información útil constituye conocimiento que conduce a mejores decisiones y se transforma en planes que impulsan acciones tácticas y estratégicas rentables. En síntesis, BI es el acto de transformar datos en bruto u operativos, o datos de cualquier fuente en información útil y significativa, para llevar a cabo

análisis de negocios y toma de decisiones [79, 125]. Este interés se ha transformado a día de hoy en un mayor número de organizaciones que utilizan este tipo de soluciones [3, 130].

Diferentes autores coinciden en esbozar los siguientes beneficios de los sistemas de BI [28, 72, 138, 37]:

- Proporciona una ayuda para la toma de decisiones en el momento correcto, en el formato correcto y a las personas adecuadas mediante la utilización de la información disponible.

- Facilita la obtención de nuevos conocimientos sobre la organización y los mercados.

- Permite a los empleados de una empresa crear y compartir información importante y filtrada para los intereses de usuarios específicos.

- Ofrece una variedad de funcionalidades como informes, análisis, uso de paneles, *scorecarding*, integración de datos, etc., ajustada a los requisitos de diferentes usuarios.

- Involucra tanto información interna (o de los sistemas transaccionales de la compañía) como datos externos a los sistemas transaccionales, así como datos actuales e históricos de la compañía, integrados y consolidados en un único repositorio.

- Aprovecha las ventajas de los datos no estructurados, como el que proporcionan las redes sociales y los objetos conectados.

- Analiza datos y genera información procesable para guiar diferentes decisiones comerciales estratégicas y tácticas de negocios. Presenta datos en informes, gráficos y tablas para permitir a los usuarios comprender la información y sacar conclusiones.

2.1.4. Nivel de madurez de un proyecto BI

Un modelo de madurez permite visualizar el estado de un proyecto o solución de BI de acuerdo a un número de niveles definidos por el mismo modelo, permitiendo así conocer la trayectoria que siguen la mayoría de organizaciones cuando evolucionan sus ambientes de BI desde sistemas tan simples como hojas de calculo hasta alternativas más complejas y completas como almacenes de datos empresariales [40, 41]. Cada nivel de madurez define el conjunto de características que debe tener la solución para estar en dicho nivel.

Eckerson, investigador del *The Data Warehouse Institute (TDWI)* [1], propuso un modelo de seis etapas para indicar la madurez de un proyecto de BI a nivel organizativo [42]. Las etapas que se proponen en este modelo son: prenatal, párvulos, niño, adolescente, adulto y sabio. Así, el valor de negocio se incrementa en la medida que la solución BI va pasando a través de cada una de las etapas. La figura 2.1 resume la curva de distribución que se tiene de las empresas que incorporan soluciones BI de acuerdo al modelo propuesto, donde la forma de campana pone de relieve que la mayoría de empresas, de acuerdo a las investigaciones del TWHI, se encuentran en las etapas de niño y adolescente, etapas en la cuales se han integrado algunas herramientas y prácticas de la BI, pero no existe una integración a nivel semántico de toda la organización.

En la Tabla 2.4 se puede apreciar las características de cada uno de los niveles de acuerdo a variables tales como arquitectura, alcance o enfoque. Se puede apreciar que a medida que se avanza en las etapas se logra un mayor nivel de integración semántica, los datos se consolidan y se aumenta el valor del negocio.

El estado ideal para una organización es el nivel sabio, donde existe una integración total y automatizada de las fuentes de datos de la organización y se obtiene el mayor nivel de autoservicio por parte de los usuarios de los datos. No

[1] *The Data Warehouse Institute* es una de las instituciones líder mundialmente en relación al aprovechamiento de los datos a nivel organizativo. Dentro de las actividades que realiza el TDWI se encuentra la emisión de informes técnicos relacionados con buenas practicas y modelos de referencia para el desarrollo de almacenes de datos y soluciones de BI. `https://tdwi.org/Home.aspx`

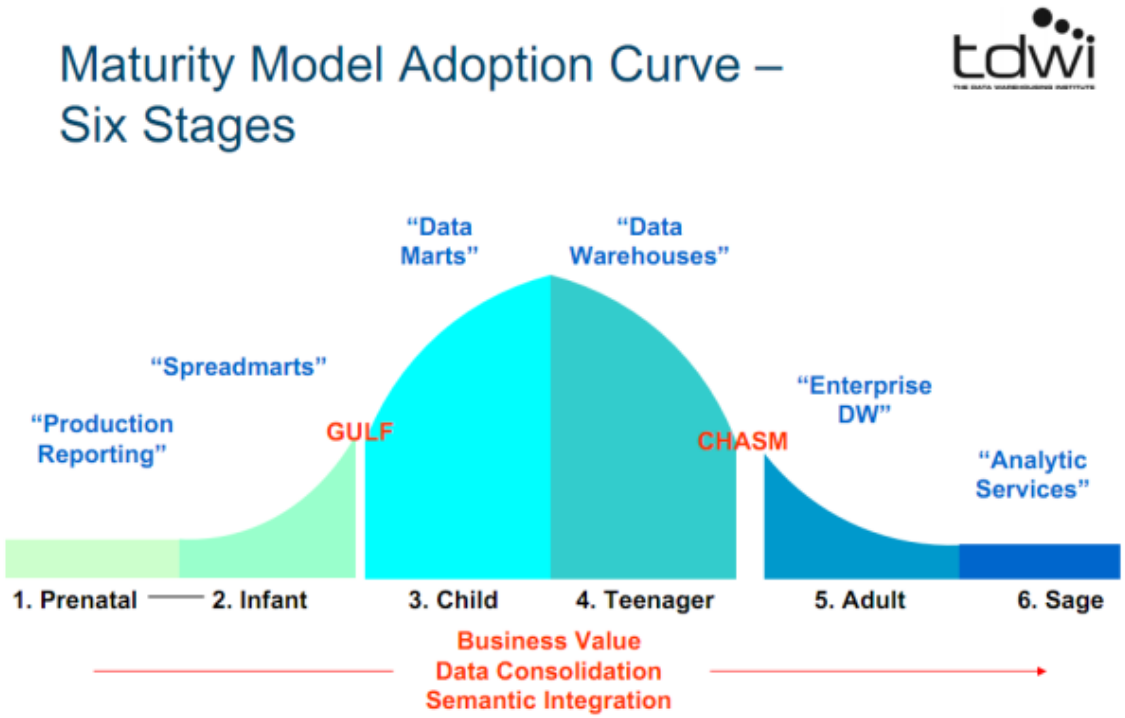


Figura 2.1: Etapas del modelo de madurez del TDWI (figura tomada de [42])

	Prenatal	**Párvulos**	**Niño**	**Adolesc.**	**Adulto**	**Sabio**
Arquitectura	Informes Administrativos	Hojas de calculo	Data marts	almacenes de datos	almacenes de datos empresariales	Servicios analíticos
Alcance	Sistema	Individual	Deptal	Divisional	Empresarial	Inter empresarial
Tipo de sistema	Financiero	Ejecutivo	Analítico	Monitoreo	Estratégico	Servicio de negocio
Analíticas	informes impresos	informes resumidos	informes interactivos	Dashboard	Scorecard en cascada	Inteligencia de negocios embebida
Usuarios	Todos	Analistas	Trabajador de conocimiento	Directores	Ejecutivos	Clientes
Enfoque	¿Qué sucedió?	¿Qué puede pasar?	¿Por qué sucedió?	¿Qué está sucediendo?	¿Qué deberíamos hacer?	¿Qué podemos ofrecer?

Tabla 2.4: Características de las etapas del modelo de madurez del TDWI (datos tomados de [42])

obstante, hay que destacar que aunque existe un grado de autoservicio en cada etapa, no es hasta la última donde se permite tanto usar información como crearla

y añadir nuevos recursos. A continuación se explica los tres niveles de autoservicio que se pueden dar en un proyecto BI:

- **Uso de la información**: es el nivel más bajo de autoservicio, los usuarios pueden acceder únicamente a la información que ya ha sido creada (informes existentes). Concretamente, el usuario solo tiene acceso a informes estándar mediante la aplicación de algunos filtros básicos. En este nivel el equipo de TIC da acceso a todos los informes que son considerados potencialmente relevantes para el trabajo del usuario. Este nivel es conveniente para usuarios ocasionales sin habilidades analíticas que solo necesitan información básica que se puede derivar fácilmente de los datos existentes. Para obtener información más profunda e individual, este nivel de autoservicio no es lo suficientemente flexible.

- **Creación de información**: en este segundo nivel de autoservicio los usuarios tienen acceso a un mayor nivel de detalle o granularidad de la información, a partir de la cual se puede generar nueva información. Esto posibilita que las organizaciones no tengan una alta dependencia del departamento TIC, a la vez que evita que estos departamentos tengan que crear un gran numero de informes que prevean todas las necesidades de información que tendrán los usuarios.

- **Creación de recursos de información**: en los dos niveles descritos anteriormente, los usuarios tienen acceso únicamente a la información que ha sido integrada con anterioridad y no pueden incorporar nuevas fuentes. Este tercer nivel extiende las funcionalidades de los otros niveles de modo que ahora los usuarios tienen la oportunidad de aprovechar de forma autónoma nuevas fuentes de datos para el análisis sin necesidad de que estas sean procesadas previamente por el departamento TIC. Es decir, pueden crean nuevos recursos de información combinando (temporalmente) las nuevas fuentes de datos con los datos corporativos (integrados con anterioridad).

Este enfoque requiere un mayor esfuerzo por parte del usuario final a la hora de identificar relaciones entre los datos y evitar el uso de datos de baja calidad.

2.2 Arquitectura

La figura 2.2 muestra la arquitectura básica de una solución BI donde se puede identificar cuatro capas, cada una con sus respectivos componentes [27, 80, 152]:

1. **Fuentes de datos**: además de las fuentes de datos necesarias y de interés para el dominio de análisis, en esta capa se ubica el componente de Extracción, Transformación y Carga (ETL, del término en inglés *Extraction Tranformation and Load*) que es el responsable de procesar los datos y depositarlos en el almacén de datos.

2. **Integración**: en esta capa se ubican las estructuras de datos necesarias tanto para el procesamiento como para la integración de los datos. El componente de mayor relevancia de esta capa es el almacén de datos que alberga los datos ya consolidados y listos para ser analizados.

3. **Procesamiento**: una vez los datos han sido limpiados, los componentes de procesamiento analítico en línea (OLAP, del término en inglés *OnLine Analytical Processing*) y minería de datos de esta capa preparan los datos para ser utilizados por el usuario final.

4. **Visualización**: en esta última capa se ubican los informes, gráficos e indicadores que permite a los usuarios tomar decisiones basadas en información veraz y oportuna.

Desde la perspectiva de usuario, una arquitectura BI se puede dividir en dos grandes partes: el *Back-End* y el *Front-End*. El *Back-End* está asociado a la recopilación y organización de datos, y el *Front-End* al análisis y visualización de los mismos. De este modo, en el *Back-End* se sitúan las fuentes de datos, el ETL, el

almacén de datos, el procesamiento analítico en línea y los modelos de minera de datos; mientras que los indicadores, gráficos e informes se situarían en la parte del *Front-End*.

En las siguientes secciones se describen cada uno de los componentes mencionados, los cuales forman parte de una arquitectura típica de una solución de Inteligencia de Negocios.

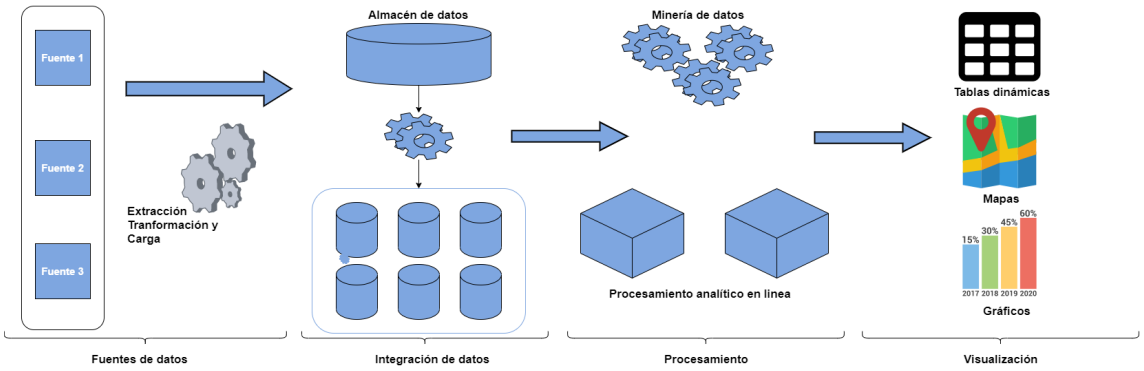


Figura 2.2: Capas de una arquitectura de Inteligencia de Negocios

2.2.1. Proceso de extracción, transformación y carga de datos (ETL)

Un proceso de BI implica identificar fuentes de datos heterogéneas que pueden ser datos operativos, datos históricos, archivos de texto plano, archivos externos o datos heredados. El proceso de ETL es una de las actividades técnicas más críticas en el desarrollo de BI ya que la integridad, uniformidad, consistencia y disponibilidad de los datos que serán utilizados por los demás componentes dependen de este componente y de su adecuada implementación.

La tarea de un diseñador de procesos de ETL, de acuerdo con Vassiliadis, Simitsis y Skiadopoulos [148], abarca las siguientes tareas: (1) analizar las fuentes de datos existentes para encontrar la semántica oculta en ellas y (2) diseñar el flujo de trabajo para extraer datos a partir de las fuentes, reparar inconsistencias de los datos, transformar los datos en un formato deseado, y, finalmente, insertarlos en el almacén de datos. El proceso de ETL, como se puede ver en la Tabla 2.5, se subdivide en tres subprocesos que se derivan a partir de su nombre. El subproceso de extracción es el responsable de leer los datos desde las fuentes,

Componente	Entradas	Operaciones	Salidas
Extracción	Fuentes de datos, sistemas transaccionales, hojas de cálculo, archivos de texto.	Selección	Datos crudos cargados en memoria principal o secundaria según lo necesitado.
Transformación	Datos crudos (cargados en memoria)	Limpieza, transformación, personalización, realización de cálculos y aplicación de funciones de agregación.	Datos formateados, estructurados y resumidos de acuerdo a las necesidades (aún en memoria)
Carga	Datos formateados, estructurados y resumidos de acuerdo a las necesidades (aún en memoria)	Inserción	Datos formateados, estructurados y resumidos con persistencia en el almacén de datos

Tabla 2.5: Subprocesos del proceso de ETL

el de transformación es el responsable de limpiar los datos, ajustarlos al formato deseado y reparar inconsistencias; y el de carga es el responsable de depositar los datos en destino.

Aunque cada uno de estos subprocesos (extracción, transformación y carga) se puede desarrollar utilizando herramientas de bajo nivel de abstracción como lenguajes de programación de propósito general, por ejemplo Python o Java, o lenguajes de propósito especifico, por ejemplo SQL, existen herramientas enfocadas específicamente a facilitar la implementación del proceso ETL. Como ejemplo de este tipo de herramientas se puede citar el Pentaho Data Integration[2] (open source) y el Oracle Data Integrator[3], las cuales permiten realizar gran parte de las tareas de manera gráfica con un enfoque de arrastrar, soltar y configurar.

2.2.2. Almacenes de datos

Los almacenes de datos (DW, del termino en inglés *Data Warehouse*) surgen como una solución a la incapacidad de los sistemas de Procesamiento de Transacciones en Línea (OLTP, por del término en inglés, *On-Line Transactional Processing*) para soportar el proceso de toma de decisiones en los niveles estratégico y administrativo. Los sistemas OLTP están pensados para soportar las operacio-

[2]http://pentaho.almacen-datos.com/kettle-spoon.html
[3]https://www.oracle.com/co/middleware/technologies/data-integrator.html

	Sistemas OLTP	Almacén de Datos
Origen de los Datos	Datos del funcionamiento diario del sistema	Datos consolidados, los datos provienen de diferentes fuentes
Propósito de almacenar los datos	Registrar operaciones del negocio. Por ejemplo: ventas.	Ayudar a planear y servir de soporte para la toma de decisiones.
Frecuencia Modificación	Frecuentemente e iniciado por los usuarios finales.	Periodos largos, e iniciados automáticamente.
Requerimientos de Espacio para el Almacenamiento	Necesita poco espacio para almacenar los datos.	Gran cantidad de espacio debido a la presencia de agregaciones y datos históricos.
Diseño Base de datos	Normalizado, por lo cual se tiene muchas tablas	Generalmente des-normalizado y con pocas tablas.

Tabla 2.6: Comparación de OLTP y almacén de datos

nes del día a día de una organización, las cuales generalmente involucran pocos datos, mientras que los almacenes de datos están enfocados para responder a las necesidades de información en la toma de decisiones, y esto involucra una mayor cantidad de datos.

Los almacenes de datos no entran en escena para remplazar a los OLTP sino para cumplir una función complementaria de análisis de datos que se denomina Fábrica de Información Corporativa [71]. En la Tabla 2.6 se muestran las diferencias que existen entre las bases de datos transaccionales concebidas con un enfoque más operativo (OLTP) y los almacenes de datos orientadas hacia el análisis.

Entre los principales beneficios de los almacenes de datos se puede citar [117, 15]:

- Reducir los tiempos de respuesta y costes de operación.
- Facilitar la toma de decisiones en los negocios a la vez que aumentar la productividad.
- Proporcionar información clave para la toma de decisiones mejorando así la calidad de las decisiones tomadas.

Bill Inmon, uno de los máximos referentes en almacenes de datos, define este término como una colección resumida y detallada de datos orientada a temas, integrada, variante en el tiempo y no volátil que se utiliza para soportar el proceso estratégico de toma de decisiones dentro de una organización [71]. Estas características atribuidas por Inmon se interpretan de la siguiente forma:

- *Orientado a temas*: los datos se ordenan según los aspectos que son de interés para la empresa. Esta característica permite responder a preguntas concretas de análisis acerca de un tema de interés, como podrían ser las ventas, las compras u otras áreas de interés de la organización.

- *Integrado*: consiste en poner los datos provenientes de diversas fuentes en un formato consistente. Esta característica permite establecer una unidad de medida común para todos los datos similares. Un almacén de datos se desarrolla integrando datos de diversas fuentes como servidores, bases de datos relacionales, archivos planos, etc., lo que obliga a mantener convenciones de nomenclatura, formato y codificación coherentes.

- *Variante en el tiempo*: los datos se interpretan como fotos de la situación de la organización, momento a momento. Todos los cambios registrados a través del tiempo quedan reflejados en el almacén de datos.

- *No volátil*: la información no se modifica ni se elimina; una vez almacenado un dato, éste se convierte en información de sólo lectura y se mantiene para futuras consultas.

Es importante destacar que existen dos visiones sobre la orientación técnica para articular un almacén de datos. Estas dos visiones tienen su origen en dos investigadores considerados los máximos referentes en área, Bill Inmon y Ralph Kimball. En la primera visión se ve al almacén de datos como una consolidación de diferentes almacenes de datos departamentales (llamados *Data Marts*); es decir, el almacén de datos no existe independiente de los *Data Marts* [71]. En la

segunda visión, el almacén de datos se ve como una fuente a partir de la cual pueden surgir los diferentes *Data Marts* [81].

2.2.2.1. Modelado del almacén de datos

Para el diseño de los almacenes de datos se suele utilizar una técnica de modelado conocida como *modelado dimensional*. A diferencia del enfoque utilizado en las bases de datos transaccionales, esta técnica maximiza la redundancia de datos para minimizar el tiempo de respuesta de las consultas y favorecer la facilidad en la exploración y uso de los datos.

El modelo dimensional se fundamenta en una relación dual hechos–dimensiones que se describe en el trabajo de Trujillo y Palomar [144]. Un **hecho** es un ítem de interés para una institución dada y se describe a través de un conjunto de atributos denominados *medidas*. Un hecho puede ser, por ejemplo, las ventas de una organización, y una medida la cantidad de productos vendidos. La **dimensión** se corresponde con la granularidad adoptada para representar los hechos y también posee atributos, llamados *atributos de dimensión*. Siguiendo con el ejemplo anterior, una dimensión para analizar las ventas puede ser el producto, el cual a su vez tiene atributos tales como nombre, marca y una categoría. Otro concepto clave del modelo son las *jerarquías*, las cuales se forman relacionando los atributos de las dimensiones, y determinan cómo se puede agregar y seleccionar las medidas para el proceso de toma de decisiones. Por ejemplo, en la dimensión producto se puede establecer una jerarquía entre categoría y producto [2, 27].

Una forma común de organizar los hechos y las dimensiones es el esquema conocido como estrella, el cual está conformado por una tabla de hechos y una o más tablas de dimensiones relacionadas a través de sus respectivas claves. La figura 2.3 muestra una tabla de hechos llamada *Ventas* rodeada de cinco tablas de dimensiones denominadas *Producto*, *Ciudad*, *Fecha*, *Vendedor* y *Usuario*. Como se puede apreciar en la figura, las dimensiones describen entidades como productos, personas, lugares y conceptos, incluso el propio tiempo.

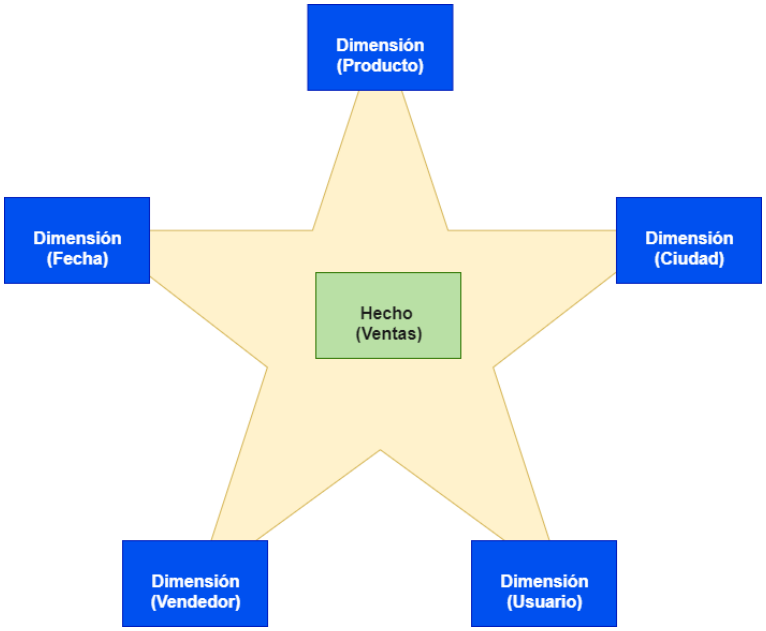


Figura 2.3: Esquema estrella para un hecho *Ventas*

La figura 2.4 muestra el esquema de la figura 2.3 implementado en las tablas de un almacén de datos. Cada uno de las puntas de la estrella representa una tabla de dimensiones y el centro de la estrella una tabla de hechos. Las tablas de hechos pueden almacenar observaciones o eventos, y pueden ser pedidos de ventas, existencias, tasas de cambio, temperaturas, etc. Una tabla de hechos contiene columnas de clave de dimensiones relacionadas con las tablas de dimensiones y columnas de medida numéricas. En la figura 2.4 la tabla de hechos de *Ventas* tiene atributos de dos tipos: (a) atributos claves como `id_producto` o el `id_vendedor` que determinan la dimensionalidad de una tabla de hechos (número de dimensiones con las que se puede analizar el hecho); y (b) medidas como la `Cantidad` y el `Valor`, que son los datos objeto de análisis. De igual forma se puede ver que cada tabla de dimensión (Producto, Cliente, Vendedor, Fecha y Ciudad) tiene sus propios atributos que pueden ser utilizados para desarrollar tareas de análisis. De este modo se puede responder a preguntas como: ¿cuántos productos de la marca X se han vendido durante el último trimestre en la ciudad de Santa Marta, Colombia?

Como conclusión, podemos decir que con el transcurrir de los años el uso de los almacenes de datos se ha extendido y hoy se habla de almacenes de datos

geo-espaciales [73], semánticos y en tiempo real, entre otras variantes [27, 10]. Adicionalmente, en cuanto a la forma de almacenamiento de los datos del almacén también han surgido diferentes alternativas. Por ejemplo, hoy es común utilizar motores de bases de datos No SQL ("NoSQL") para hacer frente a los requerimientos de escalabilidad y tolerancia a fallos, demandas que con los sistemas de bases de datos relacionales resultan extremadamente costosas.

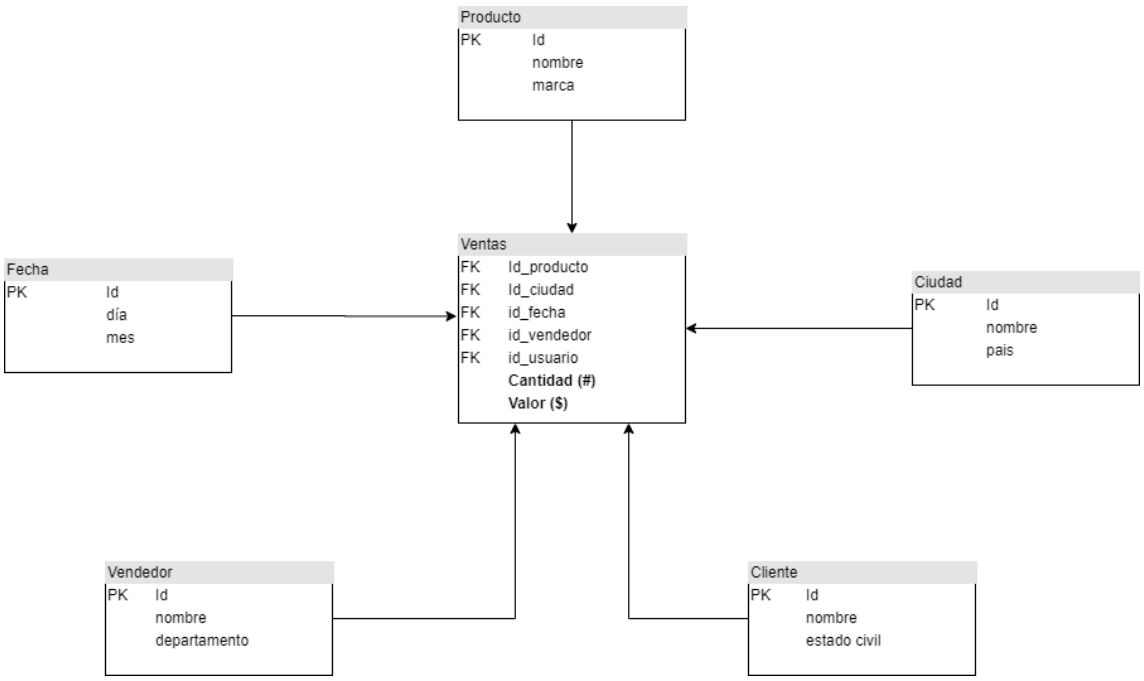


Figura 2.4: Traducción del esquema de la figura 2.3 a la representación dimensional del esquema estrella

2.2.3. Procesamiento analítico en linea (OLAP)

El análisis en línea de los datos de un almacén (OLAP) es un tipo de procesamiento de datos que se caracteriza por permitir un análisis multidimensional de la información con el objetivo de agilizar la consulta de grandes cantidades de datos [34]. El análisis OLAP se basa en modelar la información mediante el uso de hechos, medidas y dimensiones. Como se ha descrito en la sección anterior, estos elementos también están presentes en el almacén de datos, pero en OLAP se hace una representación vectorial de los mismos, mientras que en el almacén de datos su representación es a través de una base de datos relacional.

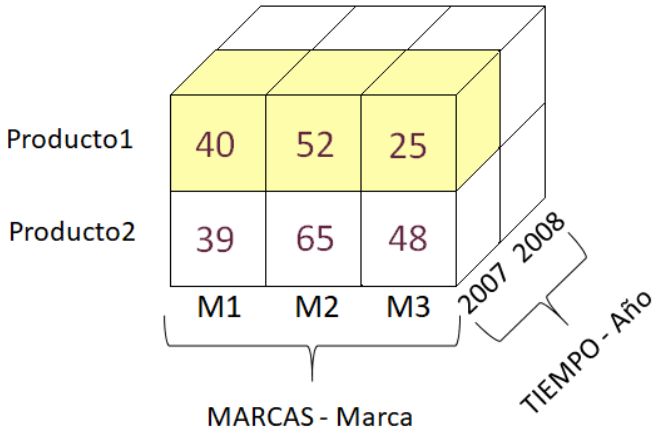


Figura 2.5: Cubo OLAP

El modelo OLAP se representa vectorialmente, esto es, los hechos se ubican lógicamente en una celda que queda en la intersección de ciertas coordenadas según el modelo de coordenadas (x, y, z,...), donde cada una de las coordenadas de la celda representa una dimensión. Esto se conoce como *análisis multidimensional* y para materializarlo en un almacén de datos se usa la correspondencia entre los elementos del modelo (hechos y coordenadas) y las partes del almacén de datos (tabla de hechos y dimensiones).

Los sistemas OLAP pueden trabajar con tres tipos de almacenamiento:

- ROLAP (Relational OLAP): son sistemas en los cuales los datos se encuentran almacenados en un almacén de datos relacional. En ROLAP se utiliza una arquitectura de tres niveles. La base de datos relacional maneja el almacenamiento de datos, el motor OLAP proporciona la funcionalidad analítica, y alguna herramienta especializada se utiliza para el nivel de presentación.

- MOLAP (multidimensional OLAP): en estos sistemas los datos se encuentran almacenados en una estructura de datos multidimensional. El sistema MOLAP utiliza una arquitectura de dos niveles: el motor de bases de datos multidimensional y el motor analítico. El primero se encarga del manejo, ac-

ceso y obtención de los datos y el segundo es el responsable de la ejecución de las consultas OLAP.

- HOLAP (Hybrid OLAP): estos sistemas mantienen los registros detallados en la base de datos relacional mientras que los datos resumidos o agregados se almacenan en una base de datos multidimensional separada. Es un híbrido de los tipos de almacenamiento ROLAP y MOLAP.

Para dar una visión del procesamiento OLAP es muy común usar la metáfora del cubo (ver figura 2.5). Un conjunto de hechos (celdas del cubo) se puede visualizar utilizando como criterio de visualización las dimensiones o ejes del cubo [98]. En la figura 2.5 el valor 40 puede representar, por ejemplo, la cantidad de ventas, el cual se describe mediante 'Producto1' de la dimensión productos, la marca 'M1' de la dimensión marcas y el año '2007' en la dimensión tiempo.

La utilización y explotación de los cubos OLAP se realiza a través de una serie de operaciones que permiten visualizar los datos que contiene el cubo desde diferentes niveles de agregación y perspectivas. A continuación se mencionan algunas de estas operaciones [27]:

***Drill-Down*:** es una operación que permite procesar los datos con un mayor nivel de detalle. Se aplica bajando por los niveles de una jerarquía definida en un cubo. Por ejemplo, en la figura 2.6 se puede apreciar un cubo con las dimensiones PRODUCTOS, MARCAS y TIEMPO. Dentro de la dimensión PRODUCTOS se tiene el atributo 'Producto' (Producto1 y Producto2); dentro de la dimensión MARCAS se tiene el atributo 'Marca' (M1, M2 y M3); dentro de la dimensión TIEMPO tenemos el atributo Año (2007). La combinación de los valores de cada uno de estos atributos cualifica una medida que se ha realizado sobre algún hecho (por ejemplo 'Ventas'); por consiguiente, el valor 40 de la figura representa la cantidad de ventas del Producto1, de la marca M1 en el año 2007. Si los atributos de una dimensión están organizados de manera jerárquica, en la parte derecha de la figura 2.6 se puede apreciar que la dimensión PRODUCTOS, además del atri-

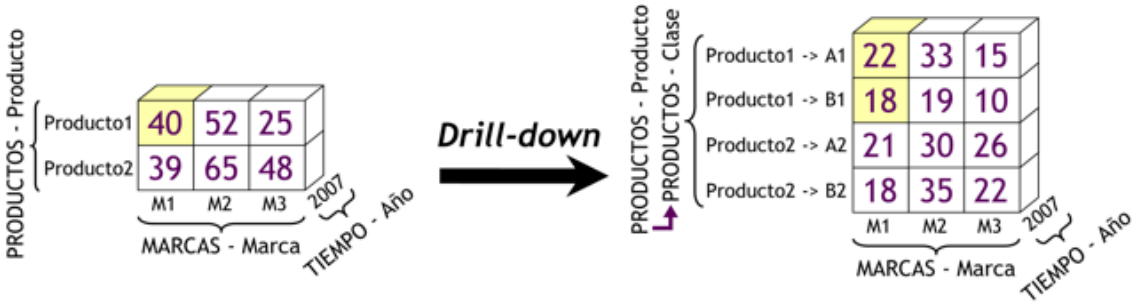


Figura 2.6: Drill down

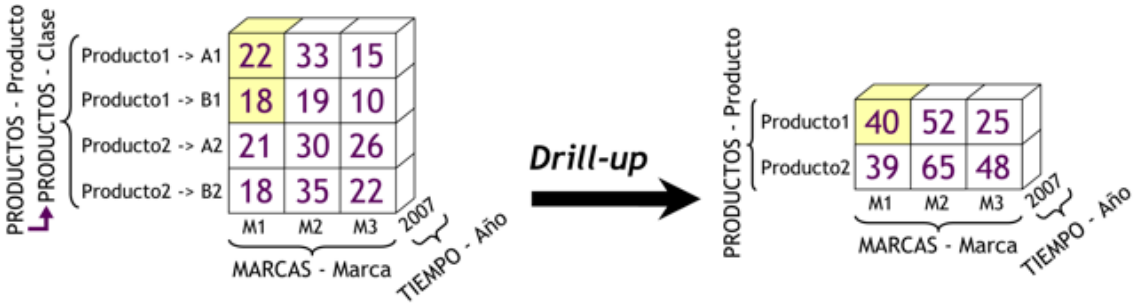


Figura 2.7: Drill up

buto 'Producto' tiene el atributo Clase (A,B) y que existe una relación jerárquica de producto y clase; es decir, los productos pueden ser de una clase. La operación *Drill-Down* permite obtener más detalle de cualquiera de los productos, especificando la clase de dichos productos. Es así como se puede ver que el valor 40 del lado izquierdo de la figura, corresponde a un 22 para el Producto1 clase A y un 18 para el mismo Producto1 clase B.

***Drill-Up*:** Esta es la operación inversa a *Drill-Down*; mientras que *Drill-Down* permite apreciar los datos a un nivel de detalle mayor, *Drill-Up* disminuye el nivel de detalle. La figura 2.7 muestra como se consolidan el valor 22 que corresponde al Producto1 de la clase A y el valor 18 que corresponde al Producto1 de la clase B en un único valor 40 que corresponde al Producto1 sin distinción de tipo.

***Drill-across*:** Esta operación en su concepción es similar a la de *Drill-Down* pero sus operandos son las dimensiones y, consecuentemente, el mayor nivel de detalle no se obtiene al descender en las jerarquías sino agregando variables independientes al indicador o variable dependiente. Esto se puede observar con mayor claridad en la figura 2.8 donde se puede apreciar en el lado izquierdo que

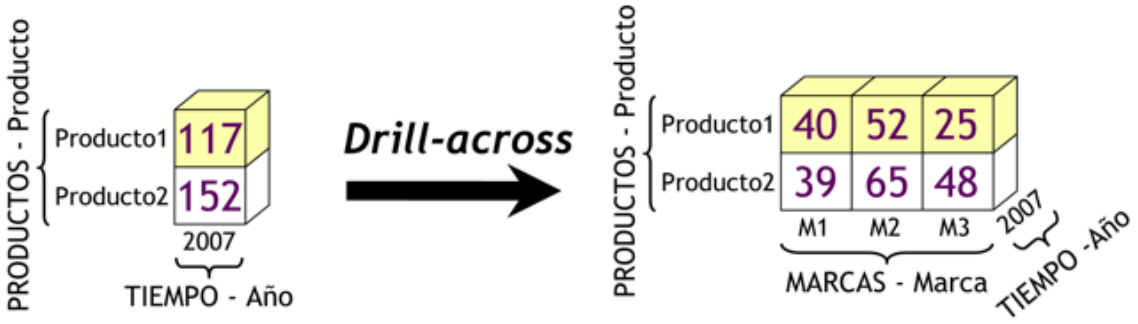


Figura 2.8: Drill across

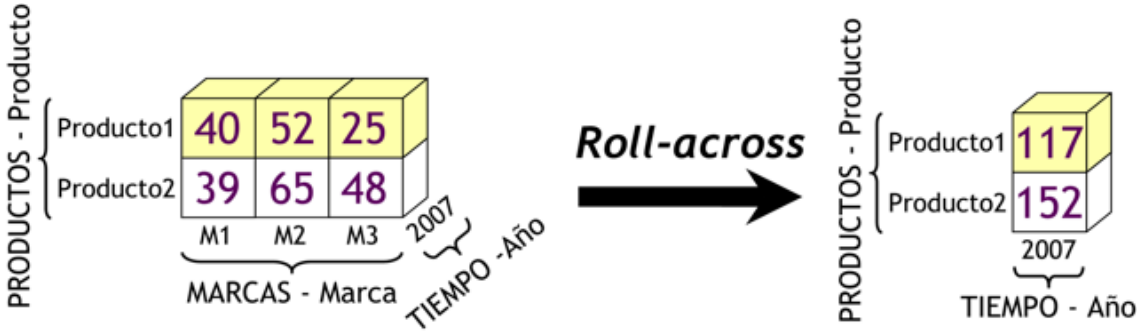


Figura 2.9: Roll across

las celdas del cubo están cualificadas por las dimensiones PRODUCTOS y TIEMPO, y de ahí que el valor 117 corresponde a las ventas del Producto1 durante el año 2007; por otro lado, en el lado derecho se añade la dimensión MARCAS por lo que las medidas son desagregadas para cada uno de los valores del atributo 'Marca'. En el ejemplo, el valor 117 de la imagen izquierda corresponde a 40 para la marca M1, 52 para la marca M2 y 25 para la marca M3 en la imagen de la derecha.

***Roll-across*:** Es la operación inversa a *Drill-across* y maneja una concepción similar a *Drill-up*. A diferencia de *Drill-across*, esta operación trabaja a un menor nivel de detalle debido a la supresión de variables, afectando directamente a la variable dependiente o indicador. Lo anteriormente expuesto se ilustra con mayor claridad en la figura 2.9.

2.2.4. Minería de datos

La minería de datos es una tecnología para resolver problemas mediante el análisis de datos existentes en el almacén de datos. La minería de datos se define

como el proceso de descubrir modelos y patrones en los datos. El proceso debe ser automático o (más usualmente) semiautomático. Los patrones o modelos descubiertos deben ser nuevos, significativos y de interés para el usuario final. La minería de datos es un término amplio que incluye varios procesos como técnicas de modelado de datos, análisis estadístico y aprendizaje automático [64].

Una de las técnicas mas representativas y distintivas de la minería de datos y, por consiguiente, de la BI es el aprendizaje automático, ya que permite manejar volúmenes creciente de datos y extraer modelos o patrones de ellos mediante un procesamiento computacional barato y potente. El aprendizaje automático se enfoca en mejorar el proceso de aprendizaje o rendimiento de los procesos algorítmicos en función de los datos.

En líneas generales, los algoritmos de aprendizaje automático se pueden clasificar en supervisados y no supervisados [127].

- Supervisado: implica que a la máquina se le enseñan modelos o busca patrones usando "ejemplos". Entonces los algoritmos supervisados intentan encontrar una función que, dadas las variables de entrada, les asigne un valor de salida adecuado luego de haber analizados las entradas y salidas de los "ejemplos". Es, generalmente, un sinónimo de clasificación. La supervisión en el aprendizaje proviene de los ejemplos etiquetados en el conjunto de datos de entrenamiento. Por ejemplo, en un trabajo de identificación de correos electrónicos no deseados, se tendría como entrada las características de los mensajes (números de caracteres, enlaces, colores del texto, etc.) y como salida si es un correo deseado o no. En el ejemplo anterior, la clasificación de cada correo como deseado o no, opera como supervisor del aprendizaje.

- No supervisado: a diferencia de los algoritmos supervisados, en los no supervisados se obtiene una inferencia a partir de los datos de entrenamiento sin necesidad de que en la entrada de datos haya una característica etique-

tada que guíe el proceso de aprendizaje. El proceso de aprendizaje no está supervisado ya que los ejemplos de entrada no están etiquetados en una clase. Es esencialmente un sinónimo de agrupamiento, y las agrupaciones se suelen utilizar para descubrir clases dentro de los datos. Por ejemplo, un método de aprendizaje no supervisado puede tomar como entrada características de un conjunto de usuarios (lugares visitados, idioma, etc.). Supongamos que encuentra dos grupos de datos. Estos grupos pueden corresponder por un lado a los usuarios que corresponden a los turistas de una ciudad y el otro grupo a los residentes. Sin embargo, dado que los datos de entrenamiento no están etiquetados, el modelo aprendido no puede decirnos el significado semántico de los grupos encontrados.

La minería de datos suele utilizarse con datos estructurados o en forma tabular donde las filas representan observaciones o individuos y las columnas características de estas observaciones. No obstante, con el auge de las redes sociales, la computación en la nube y la libertad del usuario para crear información, la mayor cantidad de datos almacenados a nivel mundial hasta la fecha corresponde a formato no estructurado (documentos, texto, vídeos, audio, etc). A tal punto que actualmente se estima que la relación de información estructurada y no estructurada almacenada electrónicamente a nivel mundial es un 20 % de información estructurada y 80 % de información no estructurado [83]. Todo ello ha llevado a que emerjan otros componentes especializados para trabajar con datos no estructurados. Así, por ejemplo, se habla de minería de texto cuando se trabaja sobre datos no estructurados o análisis de redes sociales cuando se busca, por ejemplo, hallar relación entre los integrantes de las redes sociales y sus formas de expresarse[4].

[4]Es importante aclarar que el análisis de redes sociales es un campo amplio y va más allá de hallar relación entre los integrantes de las redes sociales y sus formas de expresarse.

2.2.5. Componentes de la capa de visualización

El objetivo principal de los componentes de la capa visualización es ofrecer a través de informes, gráficos y cuadros de mando la información y el conocimiento generado en la capa de análisis [72]. La visualización y manipulación interactiva de datos bien diseñada se ha convertido en una de las principales herramientas para apoyar la identificación de patrones y la toma de decisiones por parte de los usuarios finales [43], ya que permite a los usuarios cambiar la apariencia de los informes, gráficos y cuadros de mando a través del control de algunas operaciones básicas como son la selección, e l filtro y el zoom [49].

2.2.5.1. Cuadros de mando

Uno de los componentes de visualización interactiva más populares y útiles son los cuadros o paneles de mando. Las principales características de los cuadros de mando son [46]:

- Se componen de múltiples pantallas visuales, como gráficos, vinculadas en una sola pantalla para que la información más importante se pueda monitorizar en un solo golpe de vista.
- Permite reunir todos los datos relevantes en una página e interpretarlos fácilmente. Utilizan gráficos y formularios para gerentes y empleados, convirtiéndose así en una herramienta valiosa para entornos competitivos.
- Los gerentes de las organizaciones que utilizan cuadros de mando, en lugar de necesitar tiempo para leer el contenido de informes complejos y extraer información de ellos, utilizan su tiempo para tomar decisiones simples y precisas.

2.2.5.2. Infografías

Otro componente que ha adquirido relevancia para el despliegue de información son las infografías. Una infografía se define como una visualización de datos

o ideas que intenta transmitir información compleja a una audiencia de modo que pueda utilizarse rápidamente y sea fácil de entender [79].

En correspondencia con esta amplia definición, existe hoy en día una omnipresencia de las infografías, pudiendo encontrar ejemplos en múltiples dominios, como mapas de transporte para el público, pronósticos meteorológicos o precios de acciones. Estos diferentes ejemplos demuestran como las infografías se dirigen a diferentes audiencias con diferentes propósitos. No obstante, independientemente de su uso, las infografías deben cumplir los siguientes tres objetivos:

- **Atraer:** una infografía debe involucrar al público objetivo.
- **Comprensión:** el espectador de una infografía debe entender la información fácilmente.
- **Retención:** el espectador debe recordar la información proporcionada por una infografía.

2.2.5.3. Manejo de variables cualitativas y cuantitativas

A nivel del análisis de las variables, se deben considerar dos escenarios, si la variable es de tipo cualitativo o si la variable es de tipo cuantitativo. Para las de tipo cualitativo, la tabla dinámica es la estructura de datos más utilizada debido a que proporciona una forma fácil y rápida para contar los diferentes valores de las variables. Una tabla dinámica proporciona las frecuencias de las diferentes combinación de valores de las variables, ya sea como valores absolutos o como porcentajes. De igual manera, los gráficos de barras y los gráficos de torta permiten la visualización de una variable utilizando valores absolutos o relativos.

Para la visualización de más de una variable se puede utilizar un gráfico de barras apiladas o agrupadas. En un gráfico de barra apiladas, las barras para la primera variable se dividen de acuerdo con las frecuencias de la segunda variable. En un gráfico de barras agrupadas, las barras de frecuencia para los valores

de la primera variables son mostrados de lado a lado con los valores de la segunda variable

Por otro lado, se puede obtener una descripción estándar de las variables cuantitativas calculando estadísticas de resumen como media, varianza, desviación estándar, mínimo, máximo, y cuartiles, y complementando con histogramas o diagramas de cajas y bigotes para la visualización de la distribución de frecuencia.

La aplicación de las diferentes técnicas de visualización depende esencialmente del tipo y número de variables y la complejidad de la estructura de datos. Para variables cualitativas, los gráficos de barras y los mosaicos son las técnicas más importantes. En el caso de visualización de distribuciones continuas, se puede utilizar histogramas y diagramas de cajas y bigotes.

Para el despliegue y uso de los resultados de las actividades de BI en la toma de decisiones, los resultados de los hallazgos analíticos importantes se tienen que transferir de una manera organizada y fácilmente comprensible a informes de alto nivel para no expertos. El uso de gráficos interactivos y dinámicos en estos resúmenes permite enfocarse en más detalles [132, 114, 162].

2.2.5.4. Herramientas de visualización

Las herramientas que se sitúan en la última capa de una arquitectura de BI (ver figura 2.2) se les conoce comúnmente como herramientas de BI, aunque el nombre correcto es herramientas de visualización. La razón es que los vendedores de este tipo de producto frecuentemente van dirigidos a los mercados de BI motivo por el cual utilizan términos como herramientas de Inteligencia de Negocios y herramientas de Analítica de Negocios cuando el alcance de la herramienta esta limitado a la de visualización de datos, prestándose a confusión [134].

Existen numerosas herramientas para soportar la visualización de datos. Un estudio reciente de la consultora Gartner muestra las herramientas que forman

parte de este competitivo mercado (ver figura 2.10) agrupadas en cuatro categorías [56]:

- Líderes: destacan normalmente por tener una gran cuota de mercado. Desarrollan bien su negocio en función de las características del mercado y están bien posicionados para el futuro.
- Visionarios: Son capaces de ofrecer productos innovadores. Saben hacia dónde va el mercado, pero no tienen todavía la capacidad de realizar implantaciones por su tamaño u otras circunstancias. Sería es caso de las start-ups.
- Aspirantes: Tienen buena ejecución del negocio y son capaces de dominar un gran segmento del mercado, pero no demuestran un entendimiento real de hacia dónde va éste.
- Jugadores de nichos específicos: Se enfocan con éxito en un nicho determinado, pero no adquieren una visión global ni se caracterizan por grandes innovaciones.

En la figura 2.10 también se puede apreciar como los líderes en el sector son las opciones comerciales de Microsoft, Tableau, QLIK y ThoughtSpot (propiedad de Google).

Las herramientas comerciales ofrecen una gran variedad de características para soportar fácilmente el desarrollo de componentes de visualización. Algunas características incluyen asistentes para la conexión a datos, tipologías de gráficos sofisticadas, paletas de diseño predefinidas, interacción de arrastrar y soltar, filtros, comparaciones de gráficos y manipulaciones directas de paneles. Además de estos beneficios, las herramientas comerciales pueden ser utilizadas directamente por los usuarios y apenas requieren configuración previa.

No obstante, además de las herramientas comerciales e integradas existen alternativas altamente flexibles, ligeras y fáciles de configurar dirigidas principalmente a programadores que utilizan herramientas web como HTML5, CSS, SQL,

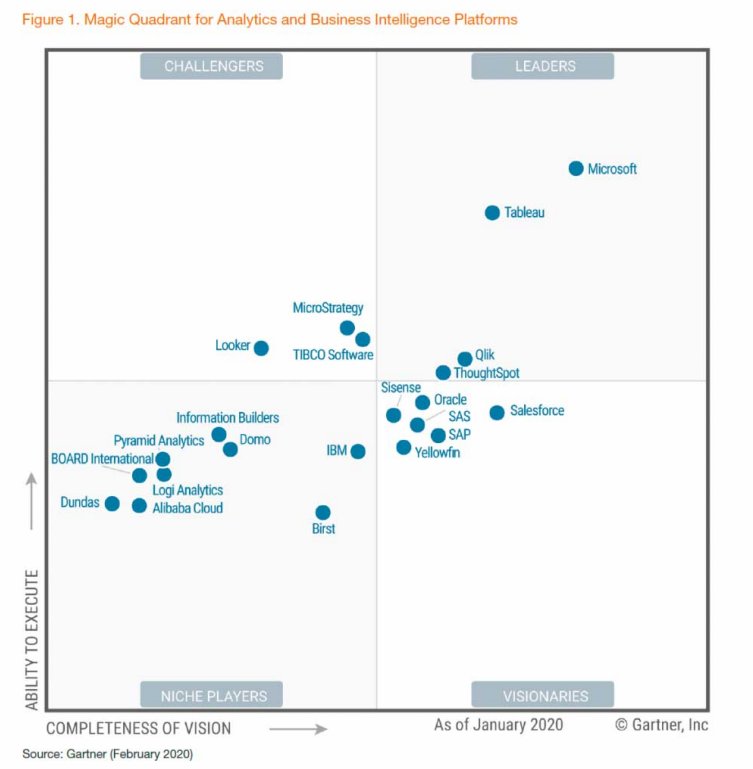


Figura 2.10: Cuadrante mágico de BI

AJAX y JavaScript para manipular los elementos de las páginas web. Algunos de los kits de herramientas configurables muy populares son D3, Prefuse y Google Chats [138].

2.3 Aplicaciones de la BI

En la literatura tanto académica como científica se han identificado varios beneficios del uso de la BI, incluyendo la optimización del trabajo operativo, mejoras en la relación con los cliente y proveedores, reducción en la redundancia de datos, facilitación de nuevos géneros de preguntas por parte de usuarios finales, mayor rentabilidad, mejor soporte para la toma de decisiones y creación de una ventaja competitiva [118, 26, 82, 123].

Uno de los sectores que hace un mayor uso de la BI es el **sector de la salud**, el cual incluye:

- Almacenes de datos, sistemas OLAP y cuadros de mandos para el seguimiento de las políticas de salud [35, 93, 17].

- Almacenes de datos espaciales que buscan el aprovechamiento de la información de los pacientes para facilitar un enfoque más eficaz de los tratamientos epidemiológicos [106, 105, 126].

- Uso de técnicas de minería de datos para crear un perfil de salud de los pacientes y comunidades para facilitar los tratamientos [153, 17, 126].

Otros contextos donde se utiliza la BI están:

- El sector comercial, donde se suele utilizar para crear perfiles de los clientes e identificar en qué gastan su dinero [51].

- El sector financiero, para integrar información interna y externa que les ayude a entender los fenómenos económicos [11, 133].

- El sector de la administración pública, como en el aprovechamiento de datos espaciales para evaluar el riesgo de erosión costera [73].

2.4 Aplicaciones de la BI en el turismo

La amplia difusión y uso de las TIC en el sector turístico facilita que información de transacciones, necesidades y comportamiento de los turistas este electrónicamente disponible. Todo este volumen de datos disponible permite que la inteligencia de negocios sea aplicada [100].

Dos hechos que han tenido un gran impacto en el incremento del desarrollo de las soluciones BI en el ámbito del turismo son:

- El incremento en el uso de la redes sociales que posibilita que haya mayor cantidad de datos disponibles creados por los turistas.

- La aparición del *Big Data* que permite almacenar y procesar grandes volúmenes de datos.

La inteligencia de negocios es de aplicación amplia en el ámbito del turismo como se evidencia en diversos trabajos de revisión de la literatura [100, 68, 67]. En estos trabajos se ocupan de extraer información de diferentes fuentes de datos y sistemas y analizar los datos con técnicas que abarcan desde visualizaciones interactivas hasta la minería de datos.

A continuación, se listan los objetivos predominantes que suelen perseguir las soluciones de inteligencia de negocios creadas en el dominio del turismo.

- Entender el comportamiento de los turistas, por ejemplo, qué lugares visita, en qué tiempo y en qué orden [102, 91].
- Saber qué piensan los turistas del destino y de sus atracciones a través de la utilización de técnicas de análisis de texto y sentimientos [86, 5, 142].
- Conocer la influencia que tiene los eventos y actividades realizadas en un destino en términos del fomento de la actividad turística [146].
- La creación de sistemas de indicadores soportados en almacenes de datos y técnicas de procesamiento analítico en linea [136, 69].
- Utilización de datos enlazados (*Linked Data*) para la recuperación de datos de diferentes fuentes para su integración en almacenes de datos para su posterior visualización [129, 1].

2.5 Trabajos previos

El uso de soluciones de Inteligencia de Negocios y fuentes de datos colaborativas se ha incrementado en la última década tanto de forma aislada como conjunta. En la literatura académica, así mismo en la científica, se han identificado varios beneficios del uso de BI, entre ellos la optimización del trabajo operativo, mejoras en la relación con clientes y proveedores, reducción de la redundancia de datos, facilitación de nuevos tipos de preguntas por parte de los usuarios finales, mayor

rentabilidad, mejor soporte de decisiones y creación de una ventaja competitiva [118, 26, 123].

Para algunos autores, la Inteligencia de Negocios es una de las facetas de los Sistemas de Soporte a la Toma de Decisiones (DSS, del término en inglés *Decision Support Systems*) [29, 12] y existen en la literatura varios ejemplos de DSS que buscan integrar diversas fuentes para facilitar el proceso de toma de decisiones. Por ejemplo, el Sistema de Información de Gestión Turística (TourMIS) [154] es un DSS financiado por la Oficina Nacional de Turismo de Austria y la Comisión Europea de Viajes y se desarrolla de acuerdo con los requisitos específicos de los administradores de turismo. TourMIS proporciona una vista integrada de varias fuentes de datos, que se pueden visualizar y analizar a través de una interfaz gráfica. TourMIS aloja datos oficiales de Eurostat y la Oficina Federal de Estadística, así como datos de turismo local y nacional proporcionados por las respectivas organizaciones turísticas, y devuelve tendencias de tasas de ocupación, número de visitantes, destinos populares, etc.

De manera similar al ejemplo anterior, la plataforma ETIHQ (*The Exposing Tourism Indicators as High Quality Linked Data*) es un DSS de turismo que se basa en TourMIS y permite visualizar y analizar indicadores estadísticos de diferentes fuentes de datos y de diferentes dominios (turismo, economía, medio ambiente) [128]. ETIHQ explota tecnologías semánticas y técnicas de minería de opiniones para procesar los datos recopilados y extraer conocimiento procesable de los repositorios. Además, muestra estadísticas de TourMIS como datos vinculados (Linked Data), lo que permite a los profesionales del turismo conectarse con otras fuentes de indicadores y explorar archivos de datos vinculados. ETIHQ experimentó dificultades en la integración de datos porque la mayoría de datos abiertos se ofrecen en diferentes formatos sintácticos, lo que implica un esfuerzo sustancial para la integración. Desde un punto de vista semántico, las dificultades provienen del uso de diferentes términos para la misma entidad, diferente granularidad geográfica o mediciones en diferentes intervalos de tiempo.

Tanto TourMIS como ETIHQ utilizan datos oficiales en gran medida. También existen intentos de integrar datos de fuentes heterogéneas [147], donde los autores presentan una aplicación de BI a la industria del turismo, específicamente, un estudio de caso de un festival gastronómico local en Tailandia. Este sistema integra datos masivos sobre productos comprados por turistas, servicios adquiridos, destinos evaluados así como datos sobre alojamiento. Posteriormente dichos datos se traducen en información significativa con el fin de que los organizadores del evento comprendan el comportamiento de los turistas, aumenten su satisfacción e impulsen los ingresos y beneficios. El marco de esta plataforma se basa en una arquitectura compuesta por sistemas de gestión de bases de datos, análisis de negocio, gestión del rendimiento empresarial, técnicas de aprendizaje automático y visualización de datos para guiar el análisis.

Por otro lado, las fuentes de datos colaborativas se han utilizando para comprender el comportamiento de los usuarios de la Web. Un claro ejemplo de esto es la utilización de Twitter para analizar el sector del turismo, complementando así a otros métodos tradicionales tales como las encuestas. Son varios los contextos en los que Twitter ha demostrado su utilidad. Un primer caso es en la identificación de turistas, donde se hace uso principalmente de la línea de tiempo del usuario para determinar los usuarios que son turistas [96]. Otro uso frecuente de Twitter es para estudiar el comportamiento de turistas, analizar aspectos cómo los puntos de concentración turistas [24] o los desplazamientos que estos realizan dentro del destino [38]. Finalmente, y no por ello menos importante, es la utilización del texto de los tuits para conocer la percepción y opinión de los turistas, es decir, para analizar el sentimiento expresado en el texto de los tuits [139].

Por su lado, OpenStreetMap (OSM) ha crecido en aceptación por parte de los usuarios y su reputación como un proyecto confiable es cada vez mayor, lo que ha llevado a que sea utilizado como proveedor de la capa cartográfica y geográfica de importantes proyectos, tanto empresariales como académicos [78, 159, 75]. A nivel empresarial, prestigiosas compañías como OpenLayer utiliza OSM para

entregar mapas fáciles de usar a través de la web. Uber, la principal compañía de transporte mediada por tecnología a nivel mundial, también utiliza OSM para la integración de información geográfica en su aplicación. OSM se ha desarrollado hasta el punto de convertirse en un vasto ecosistema de datos. Entre los usos más frecuentes de OSM se encuentra la definición de rutas de tránsito (bicicletas, personas en silla de ruedas y vehículos), permitiendo así a los usuarios de estos servicios obtener información oportuna y visual sobre las rutas y la situación del tráfico [76, 95]. OSM se emplea también en servicios de mapas para encontrar lugares específicos como casas, atracciones y restaurantes. Estos servicios permiten responder preguntas tales como: ¿Dónde está el lugar X? ¿Cuál es la distancia entre los lugares A y B? ¿Cuál es la mejor ruta entre dos puntos A y B? [14, 103, 33].

2.6 Resumen

En este capítulo se explicó en qué consiste el concepto de la Inteligencia de Negocios y la forma en como éste ha evolucionado desde su primer uso por Hans Peter Luhn en el año 1958 hasta convertirse hoy en día en un concepto sombrilla que engloba varios elementos y métodos enfocados en mejorar el proceso de toma de decisiones, utilizando sistemas basados en hechos.

De igual manera se expuso que, para cumplir con este propósito, es decir, apoyar la toma de decisiones, se hace uso de un grupo de tecnologías y componentes enfocados en diferentes tareas: los procesos de ETL para organizar el proceso de extracción, transformación y carga; el almacén de datos que permite organizar la información y dejarla en un formato consistente; los componentes de procesamiento OLAP y de minería de datos para un procesamiento más avanzado; y los componentes de visualización para hacer posible que sean accedidos y explotados tanto por usuarios expertos como por aquellos con poca destreza informática.

Finalmente, también se expuso el uso que se le ha dado a la BI en diferentes dominios como el de la salud, para analizar la expansión de las epidemias; en el

sector bancario para crear perfiles de los clientes; y en el sector del turismo para conocer las preferencias de los turistas.

CAPÍTULO 3

Metodología

El objetivo general de esta tesis es analizar la utilización de datos colaborativos procedentes de redes sociales para evaluar el sector del turismo. Para ello, se plantea un proceso metodológico compuesto de 4 fases: (1) el entendimiento del problema y las características que debe tener la solución a construir; (2) la identificación y selección de las fuentes de datos que se ajustan a las características deseadas; (3) el diseño de los componentes necesarios para la extracción, procesamiento y carga de los datos y para la estructuración, análisis y visualización de los mismos; (4) la implementación de la solución y su posterior puesta en producción. A continuación, se describen cada una de estas fases.

3.1 Entendimiento del problema

En esta primera fase se describen las propiedades que debe tener la solución a construir para cumplir el objetivo planteado, es decir, se plantean los requisitos que debe satisfacer la solución tanto a nivel funcional como no funcional. A continuación, se presenta un listado de los principales requisitos de BITOUR:

- **Fuentes de datos abiertas, colaborativas y de acceso gratuito.** Estos requisitos auto-impuestos sobre las fuentes de datos son importantes porque aportan una forma ágil y económica de conocer la opinión de las personas y su

ubicación. Nuestro objetivo es no imponer una limitación de uso de la solución a las organizaciones y entidades potencialmente interesadas.

- **Datos de tipo geográfico y social.** Los datos utilizados deben proporcionar principalmente información relativa a la ubicación de objetos (datos geográficos) y su interacción (daots sociales).

- **Recolección de datos sobre dos tipos de entidades: lugares de interés y personas.** Los datos deben proporcionar información de los lugares de interés turístico dentro del destino, como lo son los monumentos, los restaurantes y las playas; y sobre las personas que están en el destino. Dentro de la información que se debe proporcionar está la ubicación y datos complementarios para los lugares de interés y la ubicación y las opiniones para las personas.

- **Acceso automático o semi-automático a los datos.** BITOUR debe facilitar la adquisición de los datos para el trabajo de los usuarios.

- **Solución reutilizable (independiente del destino geográfico).** La solución no debe estar restringida a un destino en particular, sino que por el contrario, diversas organizaciones con diferentes necesidades de análisis puedan estudiar el destino de su interés.

- **Integración de datos geográficos y sociales.** La información de la localización geográfica de las personas se utilizará para clasificar a estas en diferentes categorías como, por ejemplo, si la persona es turista o residente dentro del destino. Esta clasificación es indispensable para las tareas posteriores de análisis.

- **Facilitación del análisis de comportamientos de personas en los lugares de interés.** La información integrada servirá para analizar comportamientos de turistas como la concentración de personas en un destino particular, lugares de interés visitados más frecuentemente, tiempo de estadía en un

lugar o la percepción de los viajeros sobre los diferentes lugares de interés que visitan.

3.2 Selección de las fuentes de datos

Como se ha mencionado anteriormente, uno de los requisitos de nuestra propuesta es que las fuentes de datos a utilizar sean colaborativas y de acceso gratuito. Por otro lado, los datos necesarios para el diseño de la solución giran en torno a dos entidades: (1) los *lugares de interés* (monumentos, hoteles, atracciones, etc.) ubicados dentro de un destino (datos geográficos); (2) *las personas* que están presentes en un destino (datos sociales). Adicionalmente, se desea extraer un conjunto de características de estas dos entidades. Específicamente, la localización geográfica de los lugares de interés y características asociadas a los mismos; y de las personas se desea recuperar su ubicación en los lugares de interés y sus opiniones acerca de estos.

La tabla 3.1 muestra el listado de fuentes de datos candidatas para obtener la información mencionada anteriormente. En las siguientes secciones se describe las razones que justifican la selección de cada una de las fuentes y se describe con mayor nivel de detalle las fuentes seleccionadas.

Tabla 3.1: Fuentes candidatas según las necesidades de información

Información a extraer	**Fuente**
1. Cartografía del destino y sus lugares de interés	OpenStreetMap, GoogleMap
2. Información de los lugares de interés	OpenStreetMap, GoogleMap, FourSquare, TripAdvisor, Airbnb, Booking
3. Ubicación de personas con respecto a los lugares de interés	Twitter, Instagram, FourSquare, Flickr
4. Opiniones de las personas sobre los lugares de interés visitados.	Twitter, TripAdvisor, Yelp

3.2.1. Cartografía del destino y lugares de interés

Para extraer las datos geográficos del destino y sus lugares de interés se barajaron dos opciones, OpenStreetMap y los mapas de Google (*Google Maps*). Ambos

sistemas son ampliamente conocidos y avalados por el público en general. Para seleccionar uno de ellos se evaluó el nivel de satisfacción a las necesidades de la tesis.

Inicialmente se consideró la libertad de utilización de los datos de estas dos plataformas. Los datos de los mapas de Google tienen derechos de autor, bien de la propia compañía Google o de las muchas organizaciones que contribuyen en la plataforma con un fin comercial. Aunque los usuarios también pueden colaborar creando contenido, no obstante, todo el contenido, creado por el usuario o no, tiene derechos de autor. En cambio, OSM es una comunidad donde todos los usuarios pueden acceder y usar los datos libremente; es decir, OSM es un proyecto abierto.

En cuando a la facilidad de acceso, ambas plataformas permiten la descarga de sus datos a través de interfaces de programación que automatizan el proceso de acceso y descarga. No obstante, OSM no establece un límite en el número de peticiones, ni restricciones en el uso de los datos descargados, siempre que se divulguen también en proyectos abiertos (como es el caso de esta tesis).

Finalmente, es clave destacar que los mapas de Google son un proyecto comercial que cuenta con el respaldo de una de las principales compañías tecnológicas de la información a nivel mundial, lo que le permite invertir en estrategias para garantizar la calidad de los datos. Mientras que OSM deja el control de calidad de los datos en mano de los usuarios. Si bien el mecanismo de control de calidad de datos OSM puede hacer pensar que sus datos son de menor calidad, diversos estudios han demostrado la confiabilidad de los mismos [45, 135].

En síntesis, por la facilidad de acceso, libertad en la utilización de los datos y la demostrada calidad de los datos se seleccionó OSM como la fuente que proporciona tanto la cartografía como la ubicación de los lugares de interés dentro del destino.

3.2.2. Información de lugares de interés dentro del destino

Los datos que proporcionan las fuentes de información geográfica como OSM se centran principalmente en las coordenadas y geometrías de los lugares, poniendo menos énfasis en información relevante como pueden ser la valoración y el precio de hoteles o los horarios de apertura y cierre de monumentos. Es por este motivo que si se desea incorporar datos complementarios para alguna categoría (subgrupo) de lugares de interés, como por ejemplo las atracciones turísticas de un destino, se debe recurrir a otras fuentes.

Este escenario fue explorado para el caso de los sitios que prestan servicios de alojamiento, más específicamente hoteles y apartamentos turísticos. Es decir, se busco una fuente que ayude a complementar la información geográfica descargada de OSM con respecto a los sitios de alojamiento con información particular de esta categoría de lugares como puede ser el precio por noche y la calificación promedio del lugar.

Para el caso de los apartamentos turísticos se seleccionó Airbnb por ser la plataforma líder a nivel mundial en este tipo de alojamiento y para los hoteles se seleccionó Tripadvisor porque posee la comunidad más activa de usuarios que contribuyen con la creación del contenido de la misma.

3.2.3. Ubicación de personas en lugares de interés

Otro requisito para el diseño de la solución que se pretende diseñar es conocer la ubicación de las personas en los lugares de interés seleccionados mediante la identificación de las coordenadas de los comentarios realizados en el destino. Para atender esta necesidad se contemplaron cuatro fuentes de datos: Twitter, Instagram, FourSquare y Flickr.

Flickr e Instagram tienen como unidad central de interacción el contenido multimedia (como las fotos). Esta información es valiosa porque permite conocer más a los usuarios y utilizar las fotos junto con los comentarios de estas como fuentes de información. No obstante, ninguna de las dos proporciona facilidades de

acceso a sus datos, ni puntos de acceso que permitan automatizar de una manera rápida y sencilla el acceso a los datos. Adicionalmente, las opiniones que se comparten en estas plataformas suelen ser principalmente positivas, por lo que se estarían omitiendo la ubicación de personas que expresan opiniones negativas.

Foursquare es una red social que permite a sus usuarios compartir su ubicación con sus contactos y publicar sus comentarios y recomendaciones de los lugares visitados. Foursquare proporciona dispone de una interfaz de programación que facilita el acceso y descarga de los datos pero el acceso gratuito se limita a la descarga de únicamente dos comentarios por cada lugar.

Twitter, por otro lado, es el líder mundial en *microblogging*. A diferencia de las otras fuentes, Twitter es una plataforma de propósito general, o lo que es lo mismo, su información no esta restringida al dominio del turismo, ni tiene que hacerse en un lugar (restaurante, estadio, etc.) en particular. Estos dos hechos representan tanto una ventaja como una desventaja. Twitter permite capturar más información pero, por otro lado, no es fácil determinar la información que es publicada por turistas. En cambio, Twitter ofrece una gran facilidad para acceder a los datos almacenados en su plataforma de manera automática y con pocas restricciones.

Para el desarrollo de este trabajo se seleccionó la plataforma Twitter ya que es la que mayor flujo de usuarios y mensajes maneja y por las grandes facilidades de acceso que proporciona a sus datos.

3.2.4. Opiniones de personas sobre lugares de interés

Finalmente, es necesario conocer la opinión de las personas sobre el destino y disponer de acceso a estos comentarios durante el periodo de estancia en el destino. Existen también diferentes plataformas que se pueden utilizar para obtener esta información, entre las que se puede citar tales Twitter, Tripadvisor o Yelp. La primera fuente, Twitter, como ya se ha mencionado, es de propósito general por lo que los comentarios, opiniones o tuits no están restringidos a un dominio particu-

lar. Mientras que Tripadvisor y Yelp centran su atención en propiciar la interacción entre personas (o usuarios de la plataforma) y lugares que prestan servicios como restaurantes, mercados, etc. En este sentido, los comentarios u opiniones que se publican en estas dos plataformas están ligados a un lugar de interés desde el mismo momento en que el usuario lo escribe debido a que para que un usuario escriba sobre un lugar antes debe seleccionarlo.

A primera vista las opciones de Tripadvisor y Yelp pueden considerarse mejores candidatas por estar enfocadas hacia servicios relacionados directamente con el turismo. Sin embargo, la desventaja frente a Twitter es que en estas plataformas existe un menor flujo de usuarios y comentarios. Por otro lado, ninguna de las dos fuentes ofrece facilidad de acceso para automatizar el proceso de descarga de datos. La única plataforma que proporciona una interfaz de programación para acceder a los comentarios (tuits) de manera totalmente automática es Twitter, por lo que fue la fuente seleccionada.

Es clave aclarar que todas las fuentes contempladas proporcionan información valiosa que merecen ser exploradas en futuras iteraciones del desarrollo de la herramienta. En la tabla 3.2 se puede apreciar un resumen de las fuentes seleccionadas y la contribución de cada una de ellas. En la primera fila se muestra que con los datos que ofrece Twitter se puede conocer la ubicación de un turista y su opinión sobre el destino.

Tabla 3.2: Relación de las fuentes

Fuente	**Información a extraer**
Twitter	- Conocer la ubicación de las personas - Poder calcular el tiempo de duración de la persona en el destino - Conocer la opinión de las personas sobre el destino en un instante específico del tiempo
OpenStreetMap	- Extraer cuáles son los lugares de interés dentro de un destino - Permitir clasificar los tipos de lugar de interés que tiene el destino.
Tripadvisor y Airbnb	- Conocer la capacidad de oferta de alojamiento para un destino - Conocer cuál es el precio manejado por cada establecimiento y la valoración que hacen las personas

En la siguiente sección se describe la razón por la que estas fuentes fueron seleccionadas, así como las limitaciones que tienen.

3.3 Descripción de las fuentes seleccionadas

En esta sección se describe en mayor detalle cada una de las fuentes de datos seleccionadas para el diseño de la solución.

3.3.1. OpenStreetMap

Esta fuente proporciona información geográfica sobre casi cualquier destino del planeta, organizada y accesible de forma automática a través de los puntos de acceso establecidos. Usar esta fuente de datos trae consigo importantes beneficios entre los que se encuentran los siguientes: todos sus datos son de acceso gratuito; sus mapas y sus datos pueden ser incrustados en cualquier aplicación sin tener que pagar por ello; sus datos están clasificados de acuerdo al uso que se le da a los objetos y cualificados con etiquetas; además de la ubicación de los objetos, proporciona información de su geometría; la calidad de sus datos ha sido puesta de manifiesto en diferentes trabajos académicos.

Sobre este último punto es importante comentar la existencia de trabajos que demuestran la validez y precisión de los datos de OSM. Por ejemplo, en [45] se presenta un análisis exploratorio que analiza la correspondencia entre la información de la base de datos de la *Corine Land Cover* sobre el uso de la tierra en Portugal y OSM. En este análisis se revisa la calidad de la clasificación de entidades poligonales de OSM contra las clases de la *Corine Land Cover* y la de su distribución espacial. Los resultados muestran una precisión de clasificación global de alrededor del 76 %. El autor de esta tesis también realizó un estudio para comparar la información turística presente en OSM con la información oficial de fuentes como la Organización Mundial del Turismo [24] y el Foro Económico Mundial (FEM) [23, 24]:

- Se evaluó la consistencia de la información contenida en el Compendio de Estadísticas Turísticas de la Organización Mundial del Turismo [156] con respecto a la información publicada en OSM, especialmente la información de lugares de alojamiento, comidas y bebidas y de agencias de viajes. Dentro de los resultados obtenidos está la alta correlación que existe entre los datos de ambas fuentes con respecto a la información del alojamiento (0.81), de los sitios de comidas y bebidas (0.87) y las agencias de viaje (0.82). [24].

- Se realizó un análisis exploratorio para comparar los datos obtenidos de OSM con los datos públicamente disponibles sobre la competitividad turística proporcionada por el FEM [155]. El análisis se realizó tomando los datos de 130 países a nivel mundial. Específicamente se estudió la representatividad de los datos del sector turístico publicados por el FEM con los de OSM. Se comprobaron ocho indicadores como el número de hoteles existentes en los destinos, el número de lugares de interés considerados patrimonio cultural y natural, etc. Los resultados mostraron que los datos de OSM proporcionan una imagen bastante precisa de las estadísticas oficiales de turismo. En general, para los los países con mayor nivel de desarrollo de la infraestructura tecnológica y de la información se observó un mayor grado de representatividad de los datos de OSM [23].

3.3.2. Twitter

En cuanto a la selección de Twitter, es clave recordar que (1) es una de las redes sociales más populares en la comunidad digital social; (2) Twitter tiene una API abierta y accesible en comparación con otras plataformas de redes sociales; (3) Twitter facilita la búsqueda y el seguimiento de conversaciones de manera automática; (4) Twitter tiene normas de *hashtag* que facilitan la recopilación, clasificación y expansión de las opiniones expresadas. Por todo lo mencionado, en Twitter se puede recuperar datos sobre eventos, noticias y opiniones de forma fácil y automatizada.

Varios trabajos han demostrado además la utilidad de Twitter para la extracción de información y conocimiento sobre patrones humanos de comportamiento. Un ejemplo de lo anterior es el trabajo presentado en [65] donde se analizan los mensajes de Twitter geolocalizados para descubrir patrones globales de movilidad humana. Basado en un conjunto de datos de casi mil millones de tuits, se estima el volumen de viajeros internacionales por país de residencia, se comparan los perfiles de movilidad de los países y se concluye que el número de visitantes estimado para diferentes países según los datos de Twitter está en consonancia con las estadísticas oficiales sobre turismo internacional. Otro trabajo que valida sus conclusiones con informes estadísticos oficiales se describe en [50], donde los tuits geolocalizados se utilizan como fuente complementaria de información para aplicaciones urbanísticas. Este trabajo presenta una técnica para determinar automáticamente los usos del suelo en un área urbana específica basada en patrones hallados en los tuits así como una técnica para identificar automáticamente los lugares con alta actividad de tuits.

A pesar de todas las ventajas que ofrece Twitter, su utilización también supone un reto. El hecho que sea una red social de propósito general, no específica para turismo (es decir, no se hace distinción entre usuarios que son turistas y usuarios que no lo son) implica el diseño e implementación de soluciones para identificar a aquellos usuarios que pueden considerarse turistas a partir de los datos recopilados desde las fuentes.

De igual manera, en Twitter existen cuentas de usuario que no corresponden a personas sino a programas informáticos que emulan el comportamiento de una persona para escribir en esta red social de manera automatizada. Estos usuarios no reales son denominados *bots*. Los *bots* añaden ruido a las tareas de análisis, es por ello que la identificación y exclusión de los mismos es clave.

3.3.3. Tripadvisor y Airbnb

Finalmente, otro requisito de este trabajo gira en torno a poder combinar la información cartográfica base de los lugares de interés del destino con fuentes de propósito específico que manejen datos sobre los servicios de alojamiento, específicamente los hoteles y los apartamentos turísticos. Las fuentes de información seleccionadas para apartamentos turísticos es Airbnb y Tripadvisor para los hoteles. La selección de la primera fuente se debe a que ésta es la plataforma líder mundial en este tipo de servicios y, adicionalmente, tiene el mayor número de apartamentos registrados en su base de datos.

En el caso de Tripadvisor, sí existen plataformas que ofrecen información similar sobre los hoteles, algunos ejemplos de estas son: Booking y Trivago, pero ninguna de estas fuentes, incluyendo *Tripadvisor*, proporciona una interfaz de programación gratuita que permita automatizar el proceso de acceso a datos. También es cierto que la presencia de hoteles en las diferentes plataformas es homogénea, por lo que el criterio que se usa para seleccionar Tripadvisor es que esta tiene la mayor cantidad de usuarios que contribuyen a la creación del contenido colaborativo de la misma.

3.4 Diseño del sistema

Con el diseño de la solución a construir, se busca dar respuesta a los siguientes interrogantes: ¿cómo integrar los datos de diferente procedencia?, ¿qué criterios utilizar para la asociación geográfica de los tuits a lugares de interés?, ¿cuáles algoritmos de identificación de bots y turistas utilizar?, ¿qué estrategia para hacer el análisis de los sentimiento expresados en los tuits usar? y ¿cuál debe ser el entorno en el que opere la solución de información que integre todas estas funcionalidades?. A continuación, se describe cómo se abordaron cada uno de estos interrogantes:

3.4.1. Integración de los datos

La primera decisión clave es cómo integrar los datos de diferente procedencia en un formato consistente de manera que permita tener un acceso fácil y rápido a los datos. Para esto existían diferentes alternativas que van desde las bases de datos relacionales tradicionales hasta las bases de datos no relacionales, pero estas alternativas no se ajustan a las necesidades puesto que están pensadas para sistemas transaccionales y no para los analíticos. La alternativa seleccionada es el uso de un almacén de datos, el cual, inicialmente fuese un lugar de integración de datos y que desde el se pudiesen desprender nuevas estructuras de datos como los Cubos OLAP, que faciliten el análisis de la información. Junto al almacén de datos vienen las correspondientes rutinas de extracción transformación y carga. Estas rutinas de extracción se ajustan a las demandas hechas por cada una de las fuentes. El detalle de este proceso se explica en el capítulo 5.

3.4.2. Asociación de los tuits a los lugares de interés

Una vez los datos han sido integrados, se hace necesario tomar otro tipo de decisiones. Si bien Twitter permite saber el punto en la tierra en el que se realiza un tuit, se debe definir qué estrategia se necesitaría para establecer que un tuit es realizado desde un lugar u otro, saber esto es clave por dos razones: primero, porque esta información resulta útil para distinguir a los turistas de los residentes en los destinos; la segunda razón, es porque esta información puede ser utilizada, por ejemplo, para analizar la concentración y gustos de los turistas. Por esto, es clave saber si un tuit es realizado desde el restaurante el "Gran Manuel" o desde el "Museo Fallero". Para esto se opta por que la asignación de un tuit a un lugar se haga utilizando los criterios de distancia y prioridad. es decir, cada uno de los tipos de objetos (museos, restaurantes, etc.) posee una prioridad que establece cuál tiene prelación sobre el otro al momento de hacer la asignación y la distancia se utilizará para decidir cuál es la máxima distancia que se permite para considerar

que un tuit es realizado desde un lugar de interés en específico. En la sección 6.1.1 se explica con más detalle este procedimiento.

Es clave destacar que los valores utilizados para ejemplificar pueden ser modificados a criterio de quien desee realizar un análisis. Los valores preestablecidos a modo de ejemplo a lo largo de este documento corresponden a los que el autor considera de mayor relevancia para el sector turístico en la ciudad de Valencia, España.

3.4.3. Identificación de bots y turistas

La identificación de turistas es una de las tareas claves dentro del proceso puesto que todos los análisis que se permiten hacer parten del supuesto que se debe conocer quiénes son turistas y quiénes son residentes, pero dado que la naturaleza de la fuente de datos seleccionada para recopilar información de los turistas (Twitter) es una fuente de propósito general, es decir, no es específica del turismo, los tuit pueden ser realizados por cualquier persona. Planteado esto, se plantea utilizar un enfoque de aprendizaje de máquina no supervisado ya que se desea describir aquellos atributos que mejor caracterizan a los turistas y no se tiene un atributo que oriente el aprendizaje. El detalle de este procedimiento se explica en la sección 6.1.3.

De igual manera, Twitter impone el reto de filtrar aquellos usuarios no reales (*bots*) que pueden añadir ruido al análisis. En esta caso, se decide considerar únicamente la proximidad espacial y temporal entre los tuits de un usuario para diferenciar a un usuario real de un bot. En la sección 6.1.2, se explica cómo se aborda la identificación de los *bots*.

3.4.4. Análisis de sentimiento

Otra tarea a realizar es definir cómo a partir del texto de los tuits (en español o en inglés) se puede analizar el sentimiento que tienen los turistas sobre su experiencia en el destino, bien sea positiva o negativa. Para esta tarea existen dos

alternativas: usar un componente existente o crear el propio modelo de clasificación. Por lo que se decide utilizar un componente ya existente puesto que permite centrar la atención en otras tareas de mayor interés y relevancia para esta tesis. La herramienta seleccionada es LIWC debido a que ha mostrado precisión entre el 0.6 y el 0.9 en trabajos similares e incorpora diccionarios de palabras tanto para el idioma inglés como para el español [36, 53]. En la sección 4.3.1.5 se explica el detalle de esta tarea.

3.4.5. Entorno de la solución

Una de las decisiones más transcendentales a tomar es definir cuál es la mejor estrategia tecnológica que permita integrar todas las decisiones tomadas hasta el momento en una sola herramienta. La respuesta a esta pregunta es una solución de inteligencia de negocios donde se unifique la extracción de datos e integración de las fuentes, técnicas de extracción de conocimiento y visualización de datos.

Finalmente, se opta por hacer la solución para ser utilizada en un entorno web debido a que se desea que dicha solución sea accedida desde diversas organizaciones interesadas a nivel global.

3.5 Implementación y despliegue de la solución

Una vez diseñada la solución que integra todos los componentes, se prosigue con su implementación. Para eso se debe decidir en qué lenguaje de programación se hará el prototipo funcional a construir, en este caso se selecciona PHP puesto que es el lenguaje que mejor maneja el autor, además que permite implementar todos los elementos diseñados de una manera rápida. No obstante, se considera que la próxima versión debe ser implementada en un lenguaje que ofrezca más facilidades en la manipulación de datos como lo es Python. A continuación se describe el resto de aspectos de la implementación y despliegue:

- **Desarrollo de bases de datos y de los procesos de ETL**. Esta fase comprende la codificación de las rutinas para extraer datos de las fuentes y el procesa-

miento de datos para derivar información que luego llenará la base de datos. Se utilizan las API proporcionadas por las fuentes de datos y se crea el código que permite acceder a ellas. Además, se codifican las rutinas para la asignación de los tuits, las rutinas para la identificación de turistas y las de cálculo de estadísticas básicas.

- **Ambiente web**. Construcción de un prototipo de aplicación web que integra todas las funcionalidades, carga de datos, procesamiento y visualización de los datos derivados.

- **Puesta en producción**. El prototipo desarrollado se despliega en un entorno de producción (servidor) con una dirección IP pública que permite el acceso desde las distintas entidades. En este servidor, todas las herramientas, bibliotecas y lenguajes utilizados por BITOUR están correctamente configurados.

Finalmente, es clave exponer cuáles son las consideraciones que se deben tener si se desea utilizar la solución en un dominio diferente al turismo y si se deben agregar nuevas fuentes de carácter específico:

- Para agregar nuevas fuentes se debe considerar que estas deben añadir información sobre alguno de los tipos de lugares de interés (por ejemplo hoteles, atracciones, etc.) cargados en la solución y se debe poder establecer una homologación entre los atributos almacenados procedentes de OSM, como el nombre y la ubicación del lugar y los de la nueva fuente.

- Para utilizar la solución en un nuevo dominio solo bastaría con indicar cuáles son las categorías de los objetos que aplican a ese nuevo dominio y modificar las reglas de asignación de tuits y agrupamiento de acuerdo a las necesidades que se tengan.

CAPÍTULO 4

Diseño de BITOUR

Tal y como se describió en el capítulo 1, el propósito de esta tesis se centra en aprovechar los datos creados de manera colaborativa para el análisis de destinos turísticos. Para alcanzar este propósito se ha diseñado una plataforma de BI que permite integrar los datos de diferentes fuentes, procesarlos y ponerlos a disposición de la comunidad para su posterior análisis y visualización.

El presente capítulo ofrece una descripción global de la plataforma construida, a la que hemos denominado *Business Intelligence platform for Tourism analysis* - BITOUR). La sección 4.1 presenta una visión general de las funcionalidades de la plataforma; la sección 4.2 detalla cómo se articulan los diferentes componentes que hacen posible su funcionamiento, es decir, la arquitectura de BITOUR. En esta segunda sección se presenta también el diseño de cada una de las capas de la arquitectura; finalmente, la sección 4.3 describe brevemente la forma en que cada una de las funcionalidades es soportada por la plataforma.

4.1 Descripción general

BITOUR es una plataforma de Inteligencia de Negocios orientada especialmente al análisis de destinos turísticos haciendo hincapié en el uso de contenido y datos creados por el usuario. El contenido que maneja BITOUR es, esencialmente, de dos tipos:

- **Espacial:** representa información relacionada con la ubicación y forma de lugares de un destino, de acuerdo a un sistemas de coordenadas geográficas.

- **Social:** representa datos de opinión que se atribuyen a un destino como puede ser texto creado en aplicaciones de *microblogging*.

Es importante destacar que aunque BITOUR está creada para llevar a cabo análisis en el dominio del turismo, la plataforma puede ser reutilizada para otros dominios siempre y cuando se respeten los dos siguientes aspectos:

- Mantener un tipo de fuente con información espacial y otro de contenido social sobre los elementos objeto de estudio del dominio en particular.

- Respetar la estructura de las capas de la plataforma de Inteligencia de Negocios que ha sido creada y la metodología de análisis seguida.

Al centrarse en las características específicas de BITOUR para el dominio de turismo, la plataforma proporciona acceso gratuito, a través de la web, a un conjunto de funcionalidades, las cuales facilitan que cualquier destino turístico interesado en analizar la percepción y comportamiento que tienen los turistas que lo visitan, pueda hacerlo. La plataforma construida exhibe las siguientes características:

- **Independiente del destino**: cualquier destino puede ser analizado siempre que se tenga acceso a los datos y fuentes de dicho destino.

- **Extensible**: nuevas fuentes de información se pueden agregar, bien sea de un dominio particular (como el turismo) o bien reemplazar las existentes con otra que proporcione información similar.

- **Acceso gratuito**: el acceso a todas las funcionalidades que ofrece BITOUR es gratuito a través de la web.

Las funcionalidades que ofrece la plataforma BITOUR son:

- BITOUR funciona para cualquier destino turístico de interés. La plataforma solicita el destino objeto de estudio y carga los datos de las fuentes a utilizar para dicho destino. Un destino puede ser cualquier zona geográfica de interés.

- La columna vertebral de la herramienta la constituyen las fuentes de datos que son independientes del dominio turístico, a saber, una con información espacial (como OpenStreetMap) y otra con información social (como Twitter). Es por esto que aunque BITOUR inicialmente funciona para el dominio del turismo, con pequeñas adaptaciones puede ser utilizada en otros dominios.

- Se pueden incorporar fuentes de un dominio específico. Por ejemplo, para el caso del turismo, y particularmente en este trabajo, se utilizan las fuentes de Tripadvisor y Airbnb.

- La herramienta permite agrupar los elementos de la fuente espacial en categorías como alojamiento, museos, restaurantes, etc.

- La herramienta asigna el contenido de la red social de opinión a un lugar dentro del destino turístico seleccionado:

    - a lugares georeferenciados y proporcionados por la fuente espacial, por ejemplo, una atracción turística, un establecimiento de comida, etc.
    - a lugares que proporcionan otras fuentes de datos especificas del dominio, como por ejemplo los hoteles presentes en Tripadvisor.

- La asignación de opiniones a un lugar se realiza de acuerdo a criterios de prioridad y proximidad, los cuales se pueden modificar dependiendo del análisis que se pretenda realizar.

- Es posible distinguir los usuarios de la red social de opinión que no representan una persona y que son sospechosos de ser una máquina que realiza post de manera automática (bots).

- A partir de los datos recolectados de la red social y un posterior procesamiento de los mismos, BITOUR permite identificar los usuarios de la red que se corresponden con turistas en el destino objeto de estudio.

- Finalmente, la plataforma permite realizar exploraciones visuales de los datos procesados.

4.2 Arquitectura

Para cumplir con las funcionalidades y especificaciones descritas en la sección 4.1 se diseñó una arquitectura de cuatro capas que sigue la misma estructura que la arquitectura genérica de BI descrita en el capítulo 2.

La arquitectura de BITOUR resume el diseño de alto nivel de la plataforma. Formalmente, la arquitectura de un sistema son las estructuras o componentes, integradas por elementos con propiedades visibles de forma externa, y las relaciones que existen entre ellos. En el caso de la Inteligencia de Negocios existe un patrón arquitectónico predominante, la arquitectura multicapa compuesta por cuatro capas: una capa donde se ubican las fuentes de datos y los procesos necesarios para su extracción, transformación y carga; una capa de integración, donde reside el almacén de datos que hospedará los datos y todas las estructuras necesarias para la unificación de estos; la capa de procesamiento donde se concentra lo relacionado con las estructuras de análisis de datos como los cubos OLAP y finalmente una capa de visualización.

La figura 4.1 muestra la arquitectura específica de BITOUR. Las siguientes secciones ilustran en detalle el funcionamiento de cada una de las capas de la plataforma BITOUR indicando también las herramientas utilizadas en el diseño de las mismas.

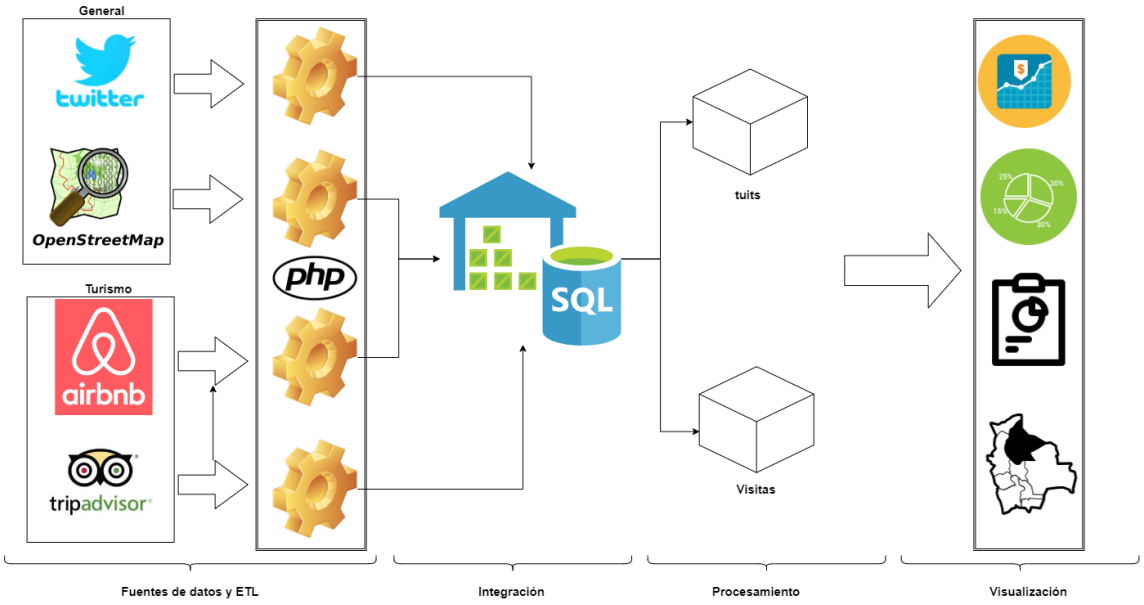


Figura 4.1: Arquitectura de BITOUR

4.2.1. Capa de fuentes de datos

Como se puede apreciar en la primera columna (de izquierda a derecha) de la figura 4.1, en el desarrollo de la presente tesis se han utilizado cuatro fuentes de datos, dos de ellas de ámbito general: OpenStreetMap (OSM) como fuente de datos espacial, y Twitter como fuente de datos de ámbito social. Estas dos fuentes de datos de propósito general proporcionan información relevante y necesaria para el análisis del sector turístico. Las razones de seleccionar OSM como fuente de datos espacial y Twitter como red social de opinión pública en BITOUR son:

- OpenStreetMap: registra la localización y geometría de millones de lugares en el mundo que son accesibles de forma gratuita a través de una interfaz de programación de aplicaciones (API, por el término en inglés, *Application Programming Interfaz*). OSM permite acceder a diferentes lugares ubicados dentro de un destino, como museos, restaurantes, atracciones, etc. Adicionalmente, OSM permite realizar consultas de datos de diversas maneras, entre las que destaca: (a) por zonas geográficas (por ejemplo, la ciudad de Valencia); y (b) por etiquetas (de la forma clave/valor) utilizadas para clasificar objetos (por ejemplo, objetos donde *tourism* = *museum*).

- Twitter: es la red de *microblogging* más utilizada a nivel mundial. Almacena la opinión, estado de ánimo y posición de millones de personas ante cualquier tipo de evento alrededor del mundo. Proporciona además acceso a sus datos de manera fácil a través de una API que permite recuperar las opiniones expresadas sobre un destino turístico. Todos estos datos se obtienen en formato JSON (del termino en inglés, *JavaScript Object Notation*).

Adicionalmente, BITOUR utiliza otras dos fuentes de datos específicas del dominio del turismo, Tripadvisor y Airbnb:

- Tripadvisor: proporciona datos sobre las facilidades y prestaciones de servicios turísticos que tiene un destino. En esta tesis se han utilizado los datos de Tripadvisor referentes a los hoteles ubicados en un destino particular. Los datos de Tripadvisor no son tan fácilmente accesibles como en OSM o Twitter puesto que no existe una API de acceso a los datos. Razón por lo que se tuvo que utilizar el *Web Scrapping*.

- Airbnb: esta fuente, similar a Tripadvisor, proporciona datos de sitios que prestan el servicio de alojamiento, principalmente con fines turísticos, pero que no pueden ser clasificados como hoteles y que muchas veces son informales. Aunque esta fuente no proporciona acceso automático a sus datos a través de una API, existen aplicaciones de terceros, como Inside Airbnb[1], que ponen los datos disponibles en formato de valores separados por comas (CSV, por el término en inglés, *Comma Separated Values*).

El proceso ETL de todas las fuentes de datos, como se observa en la figura 4.1, se implementó utilizando el lenguaje de programación PHP. En la tabla 4.1 se muestra un resumen de las cuatro fuentes de datos utilizadas en BITOUR, los datos que se extraen de cada una de ellas y el tipo de acceso. Una descripción más detallada de las fuentes y de los procesos de ETL se detalla en el el capitulo 5.

[1] `http://insideairbnb.com/`

Tabla 4.1: Resumen de la fuentes de datos utilizadas

Fuente	Datos que se extraen de la fuente	Tipo de acceso
OSM	nombre de los objetos, características en forma de etiquetas, geometría y coordenadas	API
Twitter	el texto del tuit, las coordenadas, el lenguaje asignado por Twitter, los *hashtags*, la fecha de creación del tuit y el usuario que lo realizó; ubicación e idioma del usuario	REST/JSON
Tripadvisor	el nombre de los hoteles, su ubicación, precio por noche y valoración de sus servicios.	*Web Scrapping*
Airbnb	el nombre de los hoteles, su ubicación, precio por noche y valoración de sus servicios.	CSV

4.2.2. Capa de integración

Como se observa en la figura 4.1, el objetivo de esta capa es integrar los datos procedentes de las cuatro fuentes utilizadas en un único lugar, el almacén de datos. Este almacén tiene la siguientes características;

- Se construye utilizando el enfoque propuesto por Bill Inmon [71]; es decir, primero se configura todo el modelo de datos estandarizado, y a partir de él se configuran después el resto de estructuras de análisis, bien sean almacenes de datos departamentales o cubos OLAP.
- Los datos de esta capa son básicamente de dos tipos: no espaciales y espaciales. Dentro del primer tipo se encuentra la información textual o numérica concerniente a características de objetos o entidades como puede ser el nombre de una atracción o el precio de un hotel; en el segundo tipo, se almacena la coordenadas geográficas y geométrica de un lugar.

Tal y como se indica en la figura 4.1, el almacén de datos se implementó en una base de datos SQL (o relacional). El manejo de los datos no espaciales se realizó mediante el sistema gestor de bases de datos PostgreSQL; y para manejar los datos espaciales se utilizó PostGIS, el complemento de PostgreSQL para datos espaciales. Una descripción de estas dos herramientas se ofrece a continuación.

- PostgreSQL: Es una de las opciones *open source* más interesantes en bases de datos relacionales. Es gratuito y libre, y ofrece una gran cantidad de opciones avanzadas. De hecho, es considerado el motor de base de datos (*open*

source) más utilizado en la actualidad por compañias de diferentes tamaños[2].

- PostGIS: convierte el sistema de administración de bases de datos PostgreSQL en una base de datos espacial mediante la adición de tres características: tipos de datos espaciales, índices espaciales y funciones que operan sobre ellos. PostGIS está construido sobre PostgreSQL y hereda automáticamente las características de las bases de datos empresariales, así como los estándares abiertos que implementan un Sistema de Información Geográfica dentro del motor de base de datos.

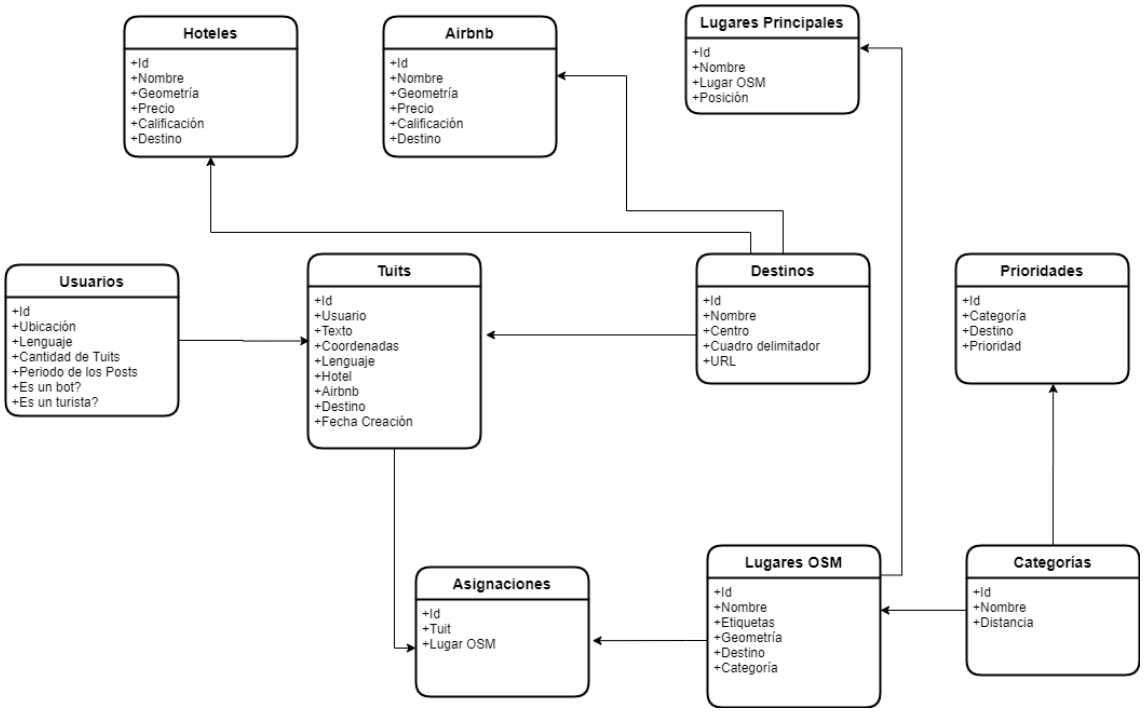


Figura 4.2: Modelo entidad relación

La figura 4.2 muestra el modelo entidad-relación del almacén de datos, el cual se compone de diez tablas. Algunas características del modelo son:

- La entidad 'destino' es transversal a las demás entidades, puesto que todos los datos y operaciones tienen lugar en el ámbito de un destino.
- Las entidades 'Hoteles', 'Lugares OSM', 'Tuits' y los sitios de 'Airbnb' almacenan datos espaciales en forma de coordenadas, geografía o geometría.

[2] `https://opensource.com/article/19/1/open-source-databases`

- Las entidades como 'Prioridades', 'Categorías', 'Asignaciones' y 'Lugares Principales' almacenan las configuraciones y resultados de operaciones realizadas por el sistema.

En la tabla 4.2 se explica brevemente el propósito de cada una de las tablas del almacén de datos; por ejemplo, la tabla 'Lugares Principales' almacena los lugares que son relevantes según el propósito del análisis que se quiera realizar.

4.2.3. Capa de procesamiento

Esta capa, como se aprecia en la figura 4.1, toma los datos integrados del almacén de datos, los procesa y reestructura para que puedan ser explotados de manera eficiente por la capa de visualización. En esta capa existen dos componentes principales: el procesamiento y los cubos OLAP:

Procesamiento. Este componente se encarga de tomar los datos, tal y como han sido integrados desde la diferentes fuentes, y realiza operaciones y cálculos sobre ellos, derivando información útil para posteriores análisis. Las tres principales tareas de este componente son:

- Asignación de tuits: permite asignar los tuits asociados a un destino a lugares particulares dentro de dicho destino, como, por ejemplo, atracciones, hoteles, restaurantes, etc.
- Identificación de bots: se utiliza para detectar los usuarios recopilados de Twitter que son un robot (una máquina programada para crear tuits).
- Identificación de turistas: la operación clave en todo el flujo es determinar los usuarios importados de Twitter que se categorizan como turistas.

Cubos OLAP. Los cubos OLAP permiten estructurar los datos de manera multidimensional para facilitar su consulta. Para este trabajo se ha seleccionado un esquema de procesamiento ROLAP de modo que los datos residen siempre en almacenes de datos departamentales. Específicamente se definen dos cubos, uno

Tabla 4.2: Descripción general del diagrama entidad relación

Tabla	Descripción
Usuarios	En esta tabla se almacena los datos relacionados con los usuarios que poden ser extraídos de twitter como lo son el id de usuario, la ubicación y el lenguaje. Además de datos calculados como la cantidad de tweets, el tiempo de estancia en el lugar, si el usuario se puede considerar un Bot y si este usuario se puede considerar un turista o no.
Hoteles	En esta tabla se almacenan los datos relacionados con los hoteles de los destinos, los cuales provienen desde Tripadvisor. Incluye datos como el nombre del hotel, precio de una noche, su calificación promedio y la ubicación.
Airbnb	Paralelo a los hoteles están los datos de los sitios que prestan los servicios de alojamiento, estos datos son procedentes del sitio de Airbnb. Acá se incluyen datos como el nombre, la ubicación, el precio de una noche y la calificación promedio.
Asignaciones	En esta tabla se almacena la asociación que se hace de cada tuit con el objeto OSM más cercano que satisface los criterios de distancia y prioridad. Se almacena un identificador del registro, el identificador del tuit y el identificador del objeto de OSM.
Tuits	En esta tabla se almacena los datos de los tuits. Entre estos se incluye el identificador del tuit, el identificador del usuario, las coordenadas del tuit, el lenguaje que Twitter asigna al tuit, el identificador del destino al cual se asociaran los tuits y el identificador del hotel y el alojamiento más cercano.
Lugares Principales	En este lugar se almacenan aquellas atracciones que son de interés para el análisis. Se almacena de ellas un identificador, el nombre de la atracción, el identificador de OSM al cual se asocia la atracción, su posición dentro del listado de todas las atracciones del destino y el identificador del destino.
Destinos	Se almacena los datos de los destinos creados en la plataforma. De cada destino se almacena un identificador del destino que posteriormente se utiliza para asociar los datos de las demás tablas a un destino en particular; el nombre del destino; el centro del destino y un cuadro delimitador de su geografía.
Categorías	En esta tabla se almacenan las categorías que son utilizadas para agrupar los objetos de OSM. teniendo atributos como el identificador de la categoría, el nombre de la categoría y la distancia que se utilizará en la asignación para considerar que un tuit está en el radio de está categoría.
Prioridades	Debido a que una categoría no necesariamente tiene la misma prioridad en todos los destino, en esta tabla se almacena las prioridades para cada categoría en un destino en particular. Por lo cual se almacena, el identificador de la prioridad, el identificador de la categoría, el identificador del destino y la prioridad.
Lugares OSM	En esta se almacenan los objetos OSM que han sido descargados por cada destino. De estos se almacena el identificador del objeto, el nombre, las etiquetas asociadas a este, la geometría, el identificador de la categoría y el destino al cual pertenecen.

que permite analizar las visitas de los turistas y otro que permite analizar los tuits realizados:

- Visitas: esta estructura está diseñada para posibilitar el análisis relacionado con los turistas, es decir, la cantidad de turistas presentes, el tiempo de su

estadía y el gasto total realizado en el destino. Para este análisis se utilizan dimensiones como las atracciones visitadas, la época del año y el tipo de atracción visitada.

- Tuits: esta estructura permite el análisis en un nivel de agregación menor, es decir, a nivel de tuit en lugar de los usuarios que lo realizan. Es así como se puede analizar la cantidad de tuits en función de si el sentimiento expresado en cada tuit es positivo, negativo o neutro; el día, mes o año en que se realizó el tuit; y los lugares desde los que se enviaron los tuits.

En el capítulo 6 se describe de manera más detallada esta capa.

4.2.4. Capa visualización

Como se muestra en la figura 4.1 esta es la última capa de la arquitectura y es la responsable de poner a disposición de los usuarios interesados toda la información disponible. Para cumplir con este propósito, BITOUR hace uso de un grupo de tecnologías que se articulan para hacer posible las diferentes tareas de análisis. Estas tecnologías son:

- *JavaScript Object Notation* para el intercambio de datos.
- *OpenLayers* para la visualización de mapas.
- HTML5 para la definición de la estructura de las paginas web.
- CSS3 y *BootStrap* para definir la apariencia de las páginas web.
- AngularJS para manejar el dinamismo de la pagina y las peticiones asíncronas al servicio de datos.
- PHP como lenguaje de programación para definir la lógica de los servicios de datos.

La figura 4.3 muestra una interacción típica de esta capa. El usuario solicita desde el navegador una dirección de un recurso; el servidor devuelve un conjunto

de datos en JSON y una pagina web (HTML5) con su estilo (CSS3); estos datos son recibidos por el navegador y a través de código en AngularJS los datos se despliegan en un mapa creado con OpenLayers. A continuación se describe cada una de las tecnologías mencionadas.

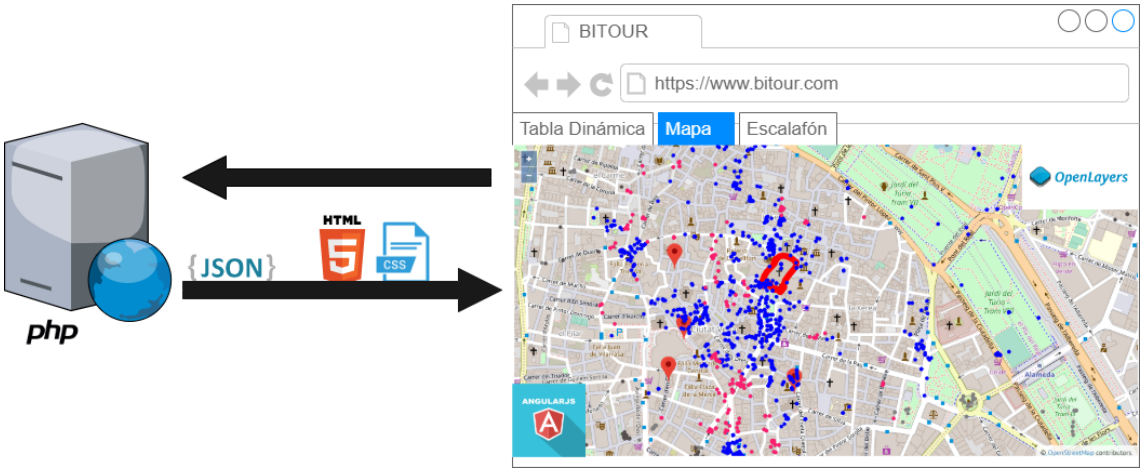


Figura 4.3: Vista general de la capa de visualización

JSON es un formato ligero de intercambio de datos. Su escritura y lectura es sencilla para los humanos y simple de interpretar y generar para las máquinas. Está basado en un subconjunto del Lenguaje de Programación JavaScript. No obstante, la popularidad y aceptación de este formato ha sido tan alta que el formato de texto que utiliza es completamente independiente del lenguaje aunque utiliza convenciones que son ampliamente conocidos por los programadores de la familia de lenguajes C, incluyendo C, C++, C#, Java, JavaScript, Perl, Python, y muchos otros. Estas propiedades hacen que JSON sea un lenguaje ideal para el intercambio de datos.

Un objeto JSON es un conjunto desordenado de pares nombre/valor. Un objeto comienza con llave de apertura ({) y termina con llave de cierre (}). Cada nombre va seguido por dos puntos (:) y los pares nombre/valor están separados por coma (,). El extracto de código 4.1 muestra ejemplo de un objeto JSON que tiene cuatro pares clave/valor:

- la primera clave define el nombre del destino (Valencia).

Fragmento de código 4.1: Representación de un objeto JSON

```
{
    destino: "Valencia",
    centro: [39.4060, -0.5081],
    geometría: [...],
    atracciones: [
                    {
                    nombre: Museo Fallero
                    tuits: [...]
                    }
                ]
}
```

- la segunda clave representa el centro geográfico del destino y se utiliza, por ejemplo, para ajustar la visualización de los mapas. El centro corresponde a un par de valores donde el primer valor corresponde a la latitud y el segundo a la longitud.
- la tercera clave representa la geometría o perímetro del destino.
- la cuarta clave un vector de atracciones ubicadas dentro del destino. Cada una de estas atracciones se representa a su vez con un objeto JSON. Cada objeto JSON dentro del vector atracciones contiene un nombre (por ejemplo Museo Fallero) y un conjunto de todos los tuits realizados alrededor de esta atracción.

OpenLayers[3] es una librería de JavaScript de código abierto para la creación de mapas interactivos en navegadores web. Esta librería facilita poner un mapa dinámico en una página web. Puede mostrar mosaicos de mapas, datos vectoriales y marcadores cargados desde cualquier fuente. OpenLayers fomenta el uso de información geográfica de todo tipo y ofrece un API para acceder a diferentes fuentes de información cartográfica en la red, tales como servicios de mapas comerciales (tipo Google Maps, Bing, Yahoo) y mapas de OpenStreetMap.

AngularJS es un framework de JavaScript de código abierto que se utiliza para crear y mantener aplicaciones web de una sola página. AngularJS fue liberado

[3] `https://openlayers.org/`

Fragmento de código 4.2: Representación de Open Layer

```
map = new ol.Map({
                target: 'map',
                layers: [
                  new ol.layer.Tile({
                    source: new ol.source.OSM()
                  }) ,alltweetsLayer ,attraccionesLayer,polygonLayer, gastronomyLayer],
                view: new ol.View({
                  center: ol.proj.fromLonLat([-0.37739, 39.46975]),
                  zoom: 12
                })
              });
```

en el año 2012 y cuenta con el soporte oficial de Google. Su objetivo es fomentar el desarrollo de aplicaciones basadas en navegador con capacidad de Modelo Vista Controlador (MVC), en un esfuerzo para hacer que el desarrollo y las pruebas sean más fáciles. Este framework es el responsable de articular los demás elementos, es decir, es el responsable de orquestar las peticiones al servidor web; recibir los datos en formato JSON; organizar la estructura HTML; organizar los estilos CSS y los datos para mostrarlos al usuario; y crear los mapas de OpenLayer para mostrar información geográfica cuando se amerite.

En el código 4.2 se muestra la instanciación de un nuevo mapa, centrado en el punto con latitud -0.37739 y longitud 39.46975 que es el centro de la ciudad de Valencia, España y luego se añaden capas de datos relacionadas con los tuits y las atracciones de la ciudad.

PivotTable fue la herramienta seleccionada para la la creación de las tablas y gráficos dinámicos. Esta es una implementación de tabla dinámica de JavaScript de código abierto (también conocida como cuadrícula dinámica, gráfico dinámico, tabla cruzada) con funcionalidad de arrastrar y soltar. El principal criterio para la selección de esta herramienta sobre otras más robustas y fáciles de usar como Talend y PowerBI fue su naturaleza gratuita. Sobre otras alternativas con acceso gratuito como dhtmlxPivot se seleccionó PivotTable porque es la más ligera y no pone restricciones en su uso.

4.3 Visión general de BITOUR

El resultado de esta tesis se materializa en una plataforma Web, BITOUR, que da vida a la arquitectura ya descrita. Para hacer posibles todas las funcionalidades, BITOUR usa las capas de la arquitectura descrita en la sección anterior y gestiona el proceso de configuración y análisis de los destinos. A modo general, el proceso soportado por BITOUR se compone de siete funcionalidades que se visualizan en la figura 4.4:

- **Definición del destino**. Como se aprecia en la figura 4.4, la primera función que ofrece BITOUR es definir el destino de interés dado que la plataforma puede operar con cualquier destino. Esta funcionalidad interactúa directamente con la capa de datos de la plataforma.

- **Carga de los datos del destino**. Permite especificar los elementos del destino que se van a cargar en la plataforma desde cada una de las fuentes. Esta funcionalidad afecta directamente la capa de integración de BITOUR.

- **Asignación de tuits a los objetos categorizados.** Los tuits obtenidos en la carga de datos se asignan a los objetos más cercanos del destino siguiendo criterios de prioridad y distancia.

- **Obtención de datos de usuarios asociados a los tuits.** A partir de los datos recopilados, por cada usuario que ha enviado tuits se calcula una serie de estadísticas y datos correspondientes al número de tuits enviados, periodo de la estancia, etc.

- **Análisis de sentimientos.** Se analiza el texto de los tuits para determinar si expresa un sentimiento positivo o negativo.

- **Identificación de turistas.** Esta función consiste en decidir los usuarios que pueden ser considerados como turistas. La funcionalidad es responsabilidad de la capa de procesamiento.

- **Análisis y visualización de datos.** Finalmente, se ponen a disposición del analista los datos descargados y procesados en las funcionalidades descritas anteriormente. El analista puede combinar y explorar los datos, crear gráficos a partir de ellos y examinar la distribución espacial de los datos resultantes con el fin de obtener información que soporte el proceso de toma de decisiones.

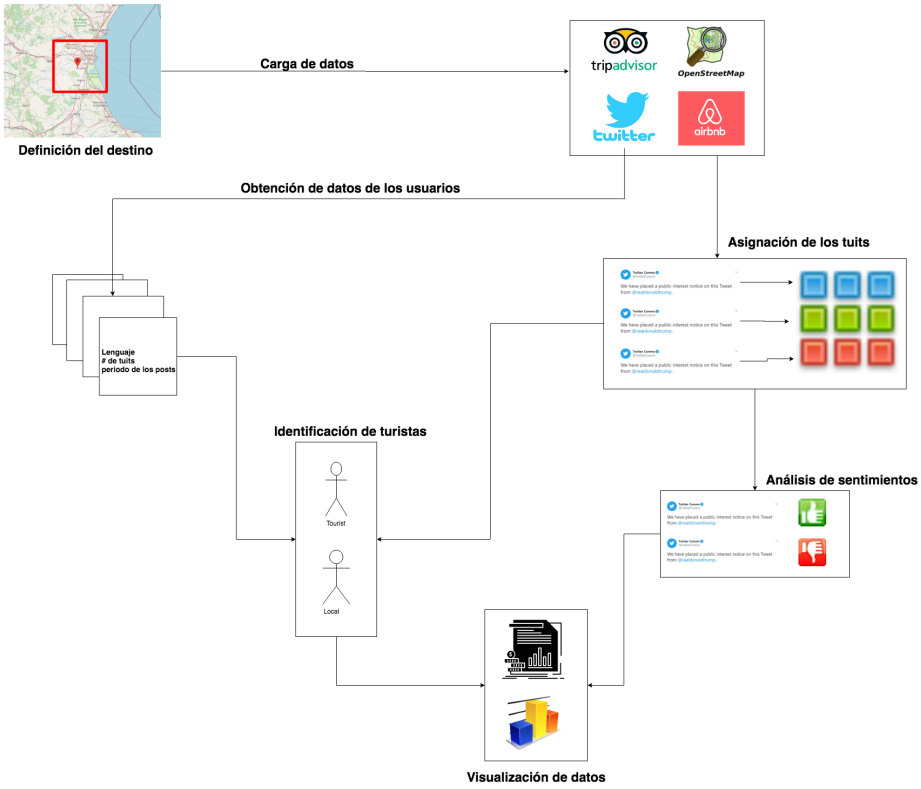


Figura 4.4: Descripción general del proceso soportado por BITOUR

Cada una de estas siete funcionalidades puede ser ejecutada por una persona de acuerdo al rol asignado (administrador o analista).

- Administrador: Este rol agrupa las funcionalidades con las que la plataforma cuenta para su configuración y para la especificación de los destinos. De este modo, la definición del destino o carga de datos, entre otras funcionalidades, solo las puede realizar un usuario bajo el rol de administrador.

- Analista: En este rol se agrupan las funcionalidades que tiene la plataforma para analizar un destino una vez se han configurado todas las variables necesarias para el análisis de datos.

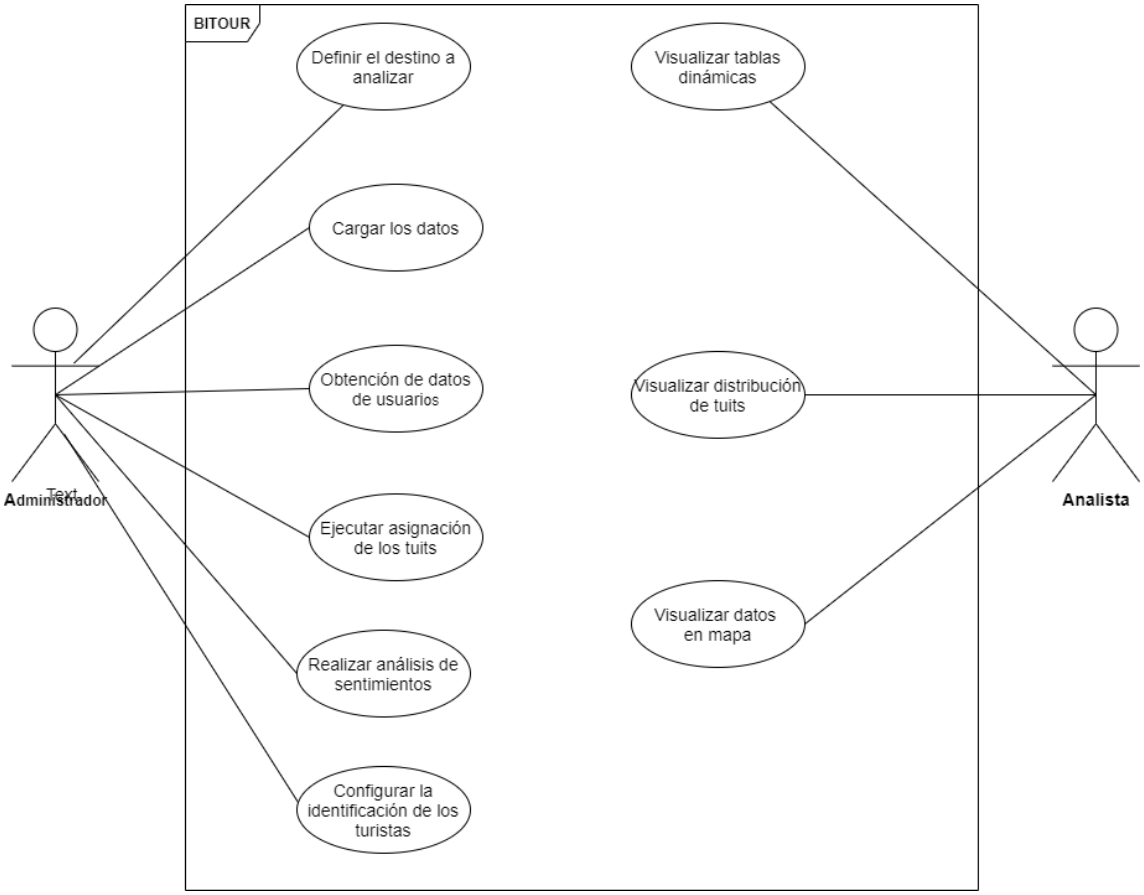


Figura 4.5: Funcionalidades para los roles de analista y administrador

La figura 4.5 muestra las funcionalidades que están bajo la responsabilidad de cada rol. El administrador tiene seis funcionalidades, todas ellas con un mismo propósito de preparar los datos del destino para su posterior análisis. Por otro lado, el analista es responsable de las funcionalidades de análisis y visualización de datos que a su vez se puede descomponer en tres tareas que son: (1) creación de tablas y gráficos dinámicos; (2) filtrado de datos en un mapa; y (3) distribución de tuits alrededor de las atracciones de un destino. Esta última funcionalidad es, quizá, la más importante porque está orientada a cumplir el propósito por el cual la plataforma fue creada, es decir, la utilización de datos colaborativos y cálculos realizados a partir de ellos para soportar el proceso de toma de decisiones en el dominio del turismo. En la sección 4.3.1 se detallan las funcionalidades del rol administrador y en la sección 4.3.2 las del rol analista.

4.3.1. Funcionalidades del administrador

Las funciones del administrador están orientadas a permitir que éste pueda establecer las configuraciones necesarias para el análisis de un destino en particular. En esta sección se detalla cada una de las funciones que puede realizar el administrador.

4.3.1.1. Definición del destino

El primer paso para realizar el análisis de un destino no incluido en la plataforma es la creación de dicho destino por parte del administrador, para lo cual se requieren los siguientes datos:

- **Nombre del destino y su descripción.** Permite definir el nombre que se asociará al destino y una breve descripción de lo que representa el destino, por ejemplo, la ciudad de Valencia, España.

- **Centro del destino.** Se debe especificar el punto central del destino o el punto que se desea utilizar para centrar todas las visualizaciones del destino. Por ejemplo, en la figura 4.6 se observa que dentro del recuadro rojo hay un marcador que indica el centro seleccionado para la ciudad de Valencia. Esto se puede realizar seleccionando el centro en el mapa o especificando los valores en los cuadros de texto `latitud` y `longitud`, donde `latitud` y `longitud` son valores decimales entre -90.0 y 90.0 y -180.0 y 180.0, respectivamente. Para el ejemplo, `latitud` es 39.4060 y `longitud` es -0.5081.

- **Cuadro delimitador**. Un paso muy importarte es delimitar geográficamente el área de análisis ya que todas las operaciones de recuperación de información se circunscribirán a dicha área. Para esto se hace uso del concepto de cuadro delimitador conocido como *bounding box*. Este cuadro delimitador es un área definida por dos longitudes y dos latitudes. Suelen seguir el formato estándar de `bbox=left, bottom, right, top`. En la figura 4.6 se representa esta área con un recuadro rojo. En BITOUR el cuadro delimitador

se especifica dibujando el cuadro con el cursor del ratón o especificando los valores de `left, bottom, right y top` en las cajas de texto creadas para este fin. En el ejemplo de la figura para el caso de Valencia, los valores son `left=-0.7635; bottom=39.1701; right= -0.20050; top= 39.5883`.

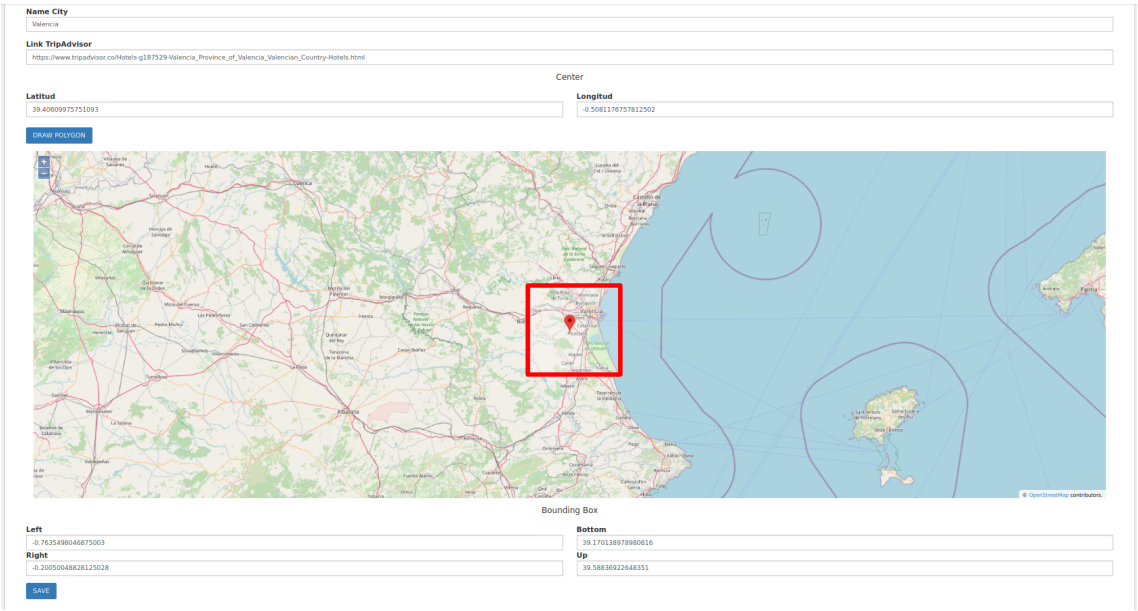


Figura 4.6: Definición del destino

4.3.1.2. Carga de los datos del destino

Las cuatro fuentes de datos integradas hasta la fecha en la plataforma BITOUR son OSM, Twitter, Tripadvisor y Airbnb. Por consiguiente, la siguiente acción de gran relevancia es cargar estos datos y asociarlos al destino. Para este fin la plataforma ofrece las siguientes funcionalidades:

- Cargar datos de OSM: la carga de datos de OSM en la plataforma se realiza utilizando unas *categorías* que se han definido previamente para agrupar los tipos de objetos de OSM que presentan características similares; por ejemplo, bajo la categoría *gastronomía* se agrupan objetos OSM que son restaurantes y bares, y en la categoría *monumentos* se agrupan objetos como sitios arqueológicos e iglesias antiguas. El administrador es el responsable de crear las categorías y definir los objetos de OSM que pertenecen a estas, utilizando

para ello los meta-datos (etiquetas) que proporciona OSM. Estas nuevas categorías creadas son independientes de las que ya pueden existir en OSM. Una vez creadas las categorías se puede proceder a cargar los objetos de OSM por cada categoría. Los detalles del procedimiento de descarga y cómo se utilizan estas categorías se explica en el capítulo 5.

- Cargar datos de Twitter: BITOUR tiene almacenados los tuits de algunas ciudades que se han descargado utilizando la API de Twitter. No obstante, teniendo en cuenta que recuperar suficientes datos es un proceso que puede tardar demasiado tiempo para ser ejecutado en línea, BITOUR tiene la opción de poder cargar directamente en la plataforma los tuits disponibles de algún destino que se hayan descargado previamente. De este modo, la carga de datos de Twitter puede hacerse en segundo plano.

- Cargar datos de Tripadvisor: esta funcionalidad permite al administrador cargar todos aquellos hoteles que están listados en Tripadvisor mediante el acceso a la página HTML de Tripadvisor para el destino seleccionado. La Figura 4.7 muestra la distribución de hoteles (puntos rojos) de la ciudad de Valencia cargados en la plataforma tras ser extraídos de la página de Tripadvisor.

- Cargar datos de Airbnb: se puede también cargar los datos de los establecimientos que prestan servicios de alojamiento y están presentes en la plataforma Airbnb. La carga de estos datos se realiza a través de un formulario que permite especificar el archivo que contiene los datos de los establecimientos. Estos datos son los proporcionados por el sitio web Airbnb Inside (más detalles de este proceso se ofrecen en el Capítulo 5).

El detalle del proceso de extracción desde cada una de estas fuentes se explica en capítulo 5.

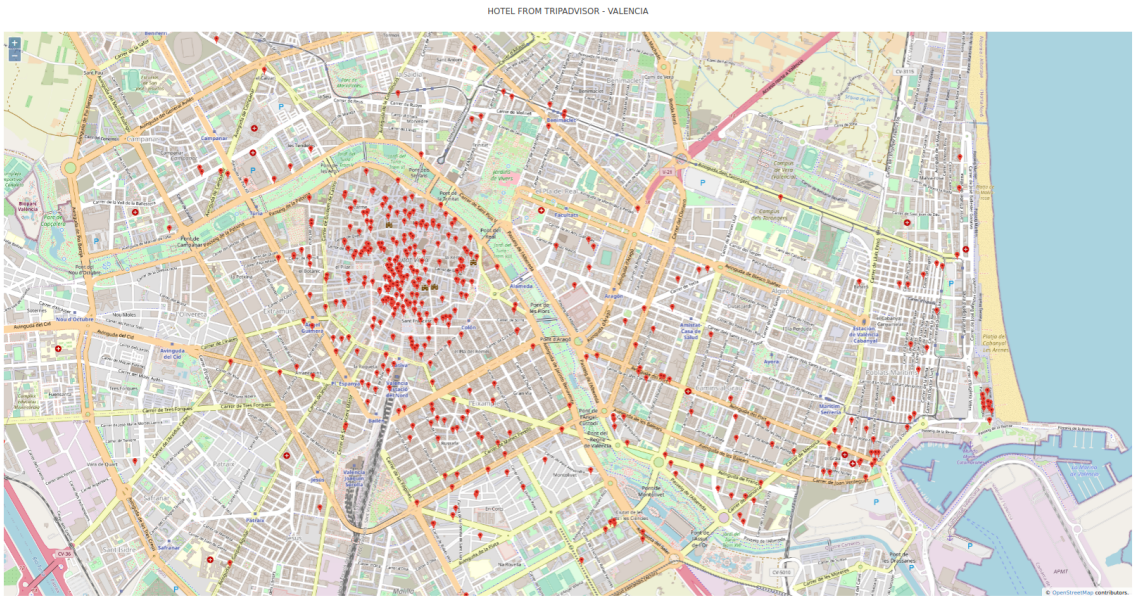


Figura 4.7: Hoteles Tripadvisor

4.3.1.3. Asignación de tuits a objetos OSM

Los tuit se asignan a objetos OSM utilizando dos criterios: la distancia entre la localización geográfica del tuit y la de los objetos OSM, y la prioridad definida en BITOUR para las categorías de agrupación de objetos OSM en función de su relevancia en el sector turístico (siendo uno la máxima prioridad). Ambos criterios, distancia y prioridad, los define el administrador para cada categoría turística. La distancia se establece en la definición de las categorías (en la sección 5.2.1.2 se describe la definición de las categorías) y las prioridades se establecen mediante el formulario de la figura 4.8.

Una vez definidos estos dos criterios, la plataforma proporciona un formulario para disparar la acción de asignar tuits a objetos OSM en segundo plano y posteriormente se notificará al correo electrónico del administrador que la asignación ha finalizado. El detalle de este proceso de asignación se explica en la sección 6.1.1.

4.3.1.4. Obtención de datos de usuarios asociados a los tuits

BITOUR realiza una serie de cálculos para extraer información de los usuarios de los tuits almacenados en la plataforma como lo son:

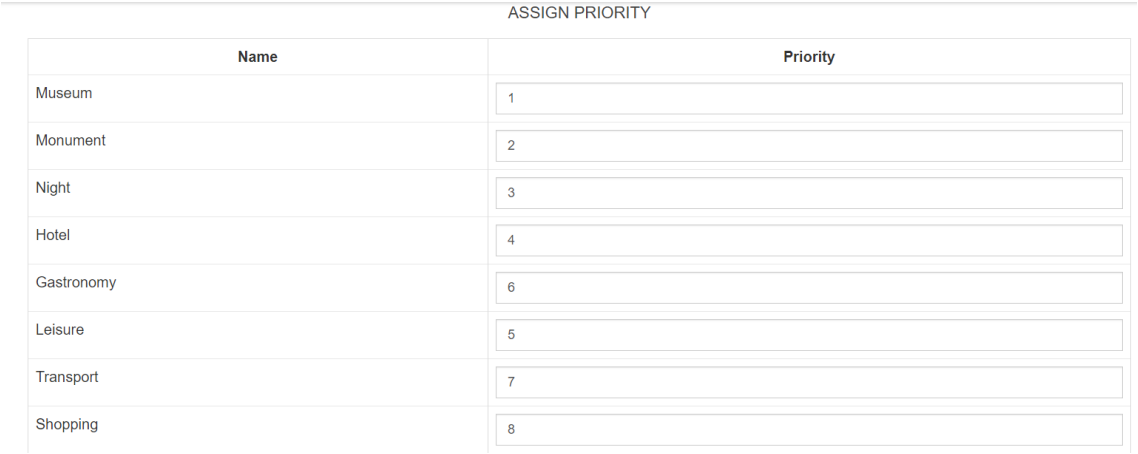


Figura 4.8: Asignar prioridades

- **Número de tuits:** para cada usuario que figura en la colección de tuits del destino, se extrae el número de tuits que ha enviado.
- **Periodo de la estancia:** la estancia del usuario en el destino se calcula con la fecha del primer y último tuit. El enfoque seguido para el cálculo de la estancia confía en los tuits realizados por un turista durante su visita a una ciudad, pero hay momentos en los que los turistas no realizan tuits, lo que puede llevar a perder precisión en el cálculo del periodo de la estancia.
- **Detección de bots:** identificar aquellos usuarios que pueden ser catalogados como no humanos, es decir, máquinas que escriben tuits de manera automática.
- **Número de tuits por categoría:** resume el número de tuits que hay asociados a un usuario por cada una de las categorías turísticas definidas en la plataforma.
- **Identificación del idioma:** la identificación del idioma del usuario se realiza siguiendo estos pasos:
  - Cuando el idioma especificado por el usuario en su cuenta de Twitter no es el inglés, se asigna como idioma el lenguaje que tiene especificado en dicha cuenta.
  - Para aquellos usuarios que tienen asignado inglés en su cuenta:

- Se asigna inglés si al menos el 75 % de los tuits están escritos en inglés.
- De lo contrario, se selecciona el idioma dominante en los textos de los tuits.

4.3.1.5. Análisis de sentimientos

Esta funcionalidad permite tomar cada uno de los tuits que han sido recolectados e identificar si la opinión que se expresa en el texto puede ser clasificada como positiva o negativa. Adicionalmente, los textos también se clasifican como religiosos, gastronómicos, entre otros. Este análisis se hace utilizando la herramienta *Linguistic Inquiry and Word Count* (LIWC), la cual permite determinar el grado en que las personas utilizan palabras que connotan emociones positivas o negativas, auto-referencias, palabras extensas o palabras relacionadas con temas de sexo, comida o religión.

4.3.1.6. Identificación de turistas

Una vez se tiene la información de usuario descrita en las secciones anteriores, el administrador procede a ejecutar la identificación de turistas. Este proceso se realiza a partir de los datos descritos y con técnicas de agrupamiento que se describirán en el capítulo 6. Al igual que en los casos anteriores, este proceso no se realiza en línea sino que se ejecuta en segundo plano y se confirma vía correo electrónico una vez el proceso ha terminado.

4.3.2. Funcionalidades del analista

En este rol se agrupa el conjunto de funcionalidades que permite al analista utilizar la configuración realizada por el administrador y los datos cargados para soportar el proceso de toma de decisiones. Del análisis y visualización de datos se desprenden tres tareas: (1) la combinación de datos a través de la creación de tablas y gráficos dinámicos; (2) la exploración espacial de los mismos, a través del filtrado en mapas según los valores de diferentes variables; y finalmente, (3) la concentración de datos en zonas de interés mediante la distribución

de los tuits alrededor de las atracciones. Es así como se puede examinar la distribución de turistas, su estadía en el destino, la concentración alrededor de las diferentes atracciones, entre otras preguntas de interés. Para lo anterior, se pone a disposición de los interesados la posibilidad de visualizar los datos de tres formas principalmente.

4.3.2.1. Tablas y gráficos dinámicos

En esta tarea se brinda la oportunidad al analista de poder acceder y combinar la información disponible en el almacén de datos ubicado en la capa de integración (Segunda capa), buscando así, enriquecer la visualización de los datos con el fin de analizarlos posteriormente. Específicamente la combinación se puede realizar utilizando tablas y gráficos dinámicos que combinan las variables disponibles de manera similar al PowerPivot de Excel, convirtiendo el conjunto de datos en una tabla o gráfico. En esta tarea, se pueden visualizar los datos en diferentes formatos en una verdadera interfaz de usuario de arrastrar y soltar en 2D.

4.3.2.2. Filtros y visualización en mapa

Permite explorar la distribución espacial de los tuits alrededor del destino. Para ello se ofrece la interfaz de la figura 4.9. En esta interfaz se muestra el mapa del destino bajo análisis con los tuits realizados en él como puntos de color azul. Los datos que se muestran en el mapa se filtran a partir de un conjunto de variables. Dentro de las posibles variables que permiten filtrarse en el mapa, se encuentran las siguientes: Tiempo, que permite saber cuándo se realizó el tuit; el lugar, para especificar desde cuál lugar se realizó; y el sentimiento, para conocer la percepción del usuario expresada en el tuit.

A modo de ejemplo la figura 4.9 representa los tuits de la ciudad de Valencia realizados por turistas alojados en establecimientos de AIRBNB. Es clave destacar que los datos representados en la figura anterior dependen de la información guardada en el almacén de datos y que las variables que pueden ser utilizadas pa-

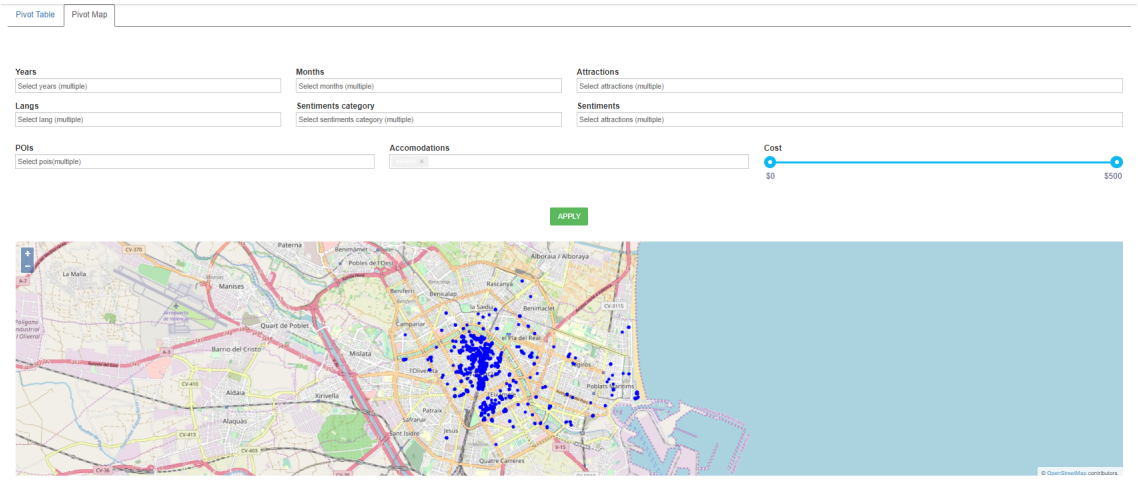


Figura 4.9: Análisis espacial

ra los filtros son configurables, es decir, el administrador puede especificar cuáles aparecerán en pantalla.

4.3.2.3. Distribución de tuits por atracciones

Esta funcionalidad permite hacer un análisis de la concentración de turistas alrededor de una serie de sitios considerados de interés turístico. La plataforma permite interactuar con el mapa del destino para seleccionar un sitio turístico en el mapa, y muestra las estadísticas de los tuits publicados dentro de un radio previamente definido, 500 metros por defecto, pero que puede ser editado en cualquier momento. También muestra estadísticas generales sobre tuits ubicados en las atracciones y lugares de gastronomía en cada uno de los destinos.

La figura 4.10 muestra a modo de ejemplo la distribución de tuits alrededor de las diez atracciones que se han catalogado como las más populares de la ciudad de Valencia.

4.4 Resumen

En este capítulo se ha abordado el diseño de BITOUR siguiendo una metodología de BI de cuatro capas que utiliza como insumo fuentes de datos de naturaleza colaborativa. La capa de datos está formada por cuatro fuentes que alimentan la plataforma: OpenStreetMap, Twitter, Tripadvisor y Airbnb. Las fuentes de

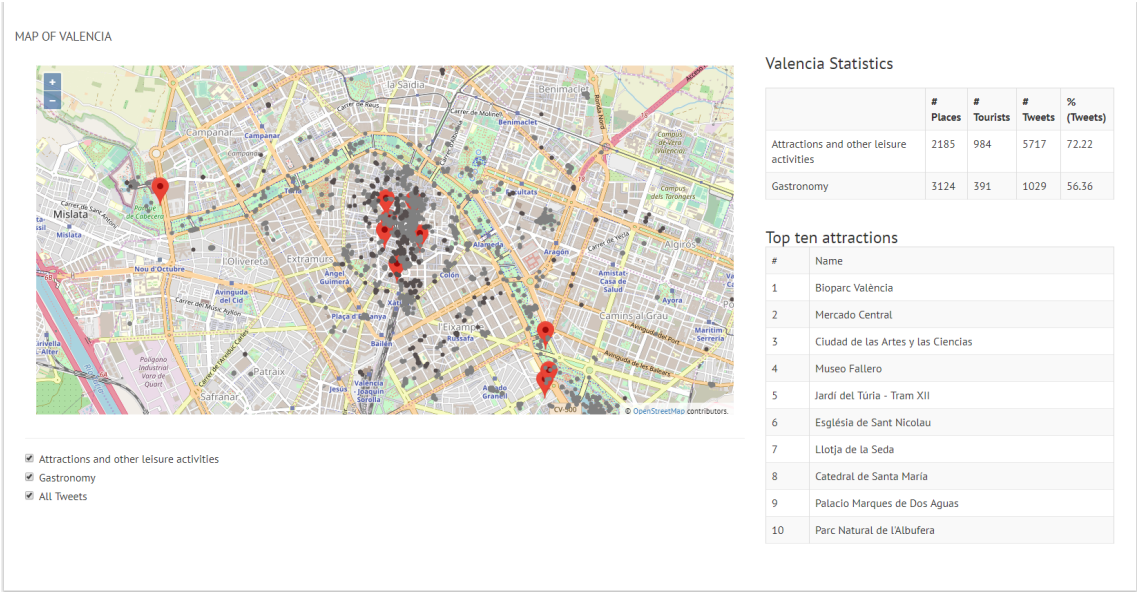


Figura 4.10: Vista general de la distribución de los tuits

esta capa son accedidas por el proceso de ETL responsable de extraer los datos, transformarlos y depositarlos en el almacén de datos (que consta de diez tablas encargadas de organizar la información) de la capa de integración. Luego, en la capa de procesamiento están los cubos OLAP, responsables de estructurar multidimensionalmente los datos para facilitar su análisis y esta capa también es responsable de ejecutar los procedimientos de minería de datos que permiten, por ejemplo, la identificación de los turistas. Finalmente, la capa de visualización es donde residen las rutinas para acceder vía web a todos los datos ya procesados.

Adicionalmente se presentó una visión general de las funcionalidades de BITOUR. La sección 4.3 describe cada una de las siete funcionalidades y cómo estas se distribuyen en los dos roles predefinidos en BITOUR: el administrador, quien accede a todas las funcionalidades relacionadas con la creación y configuración de los destinos que se desean analizar; y el analista, quien dispone, para soportar el proceso de toma de decisiones, de tareas de visualización.

CAPÍTULO 5

Fuentes de datos

En el presente capítulo se presentan las cuatro fuentes de datos utilizadas en BITOUR, haciendo especial énfasis en su estructura, el uso que se ha hecho de estas fuentes y el proceso de ETL para cada una de ellas. El capítulo comienza con una explicación de la generación colaborativa de contenidos y posteriormente se describe cada una de las fuentes.

5.1 Fuentes de datos colaborativos

Al inicio de la Web, el modelo para la creación de contenido estaba pensado para que los propietarios de los sitios publicados en la web crearan todo el contenido que deseaban compartir con los usuarios. La interactividad entre los usuarios y las páginas contenidas en estos sitios era baja y la web se concebía como un recurso que se utilizaba para consultar información, es decir, de una manera unidireccional. Sin embargo, la evolución a la Web 2.0 trae consigo varios cambios en la concepción de la creación y compartición de información [16].

El concepto Web 2.0 se acuñó en 2003 y se refiere a una nueva forma de concebir al usuario, pasando de ser un actor pasivo que recibe información a un actor proactivo que crea contenido y relaciones con otros usuarios. Esto implica la creación de sitios web dinámicos en los que el usuario puede interactuar, generar contenido, o formar parte de comunidades virtuales. Este cambio de paradigma

supuso el auge de los *blogs*, las redes sociales y otras herramientas relacionadas [9, 110]. Algunos ejemplos de páginas Web 2.0 son las redes sociales, las wiki, las páginas de ventas por Internet u otros proyectos colaborativos en los que los usuarios puede generar contenido y no simplemente consumirlo [44, 89].

Se considera que la transición de la Web 1.0 a la Web 2.0 es uno de los principales impulsores que ha traído consigo la necesidad del *Big Data* [150]. Como consecuencia del fuerte aumento en la Web 2.0 en la última década, la forma de comunicación de las personas con servicios y aplicaciones ha cambiado, y las redes sociales han emergido como fuente de información y como un medio de comunicación [81].

5.1.1. Redes sociales

Existe en la actualidad un alto número de usuarios que dedica gran parte de su tiempo a interactuar con otras personas vía redes sociales, así como a expresar sus opiniones y pensamientos acerca de varios temas. Hoy los usuarios juegan un rol muy importante en la creación de grandes cantidades de datos en el día a día sobre una gran variedad de temas (lanzamiento de un nuevo teléfono, su último viaje, un nuevo restaurante) [115], compartiendo sus puntos de vista y opiniones cómodamente en las redes sociales. Como resultado, muchas empresas desarrollan una "Inteligencia Social" basada en información extraída de las redes sociales [112].

Entre los diferentes tipos de web colaborativas más relevantes se puede citar [115]:

- Redes sociales como Pinterest, Foursquare o Twitter: los usuarios crean su propia red personal e intercambien información con base en sus intereses personales.
- Proyectos colaborativos como Wikipedia o OpenStreetMap: las personas se pueden suscribir a sitios colaborativos y crear o modificar el contenido de la página.

- Comunidades virtuales como juegos online: los usuarios crean cuentas y se unen a comunidades de su interés.
- Sitios de crowdfunding como Kickstarter: los participantes crean una página para describir su idea para un proyecto y solicitan financiación del público.
- Sitios de comercios electrónico con características de contenido generado por el usuario como Amazon: permiten a los usuarios hacer revisiones y recomendaciones de los elementos allí ofrecidos.
- Sitios para compartir contenido como YouTube: los usuarios suben y comparten contenido con otros usuarios, usualmente gratis.

Los gerentes de marketing han descubierto los beneficios de escuchar en las redes sociales así como aprovechar las oportunidades que estas ofrecen para interactuar con los clientes [111, 13, 99]. El análisis de redes sociales se ve hoy en día como un elemento de apoyo que ayuda a empresas y sociedad a comprender las necesidades y motivaciones de los ciudadanos. Comprender la mente del ciudadano permite a las empresas personalizar un mensaje general en una forma dinámica de satisfacer las necesidades de la mayoría. A la larga, captar intereses y monitorizar los sentimientos del público sobre un tema puede llevar a modelos de planificación y operaciones logísticas más eficientes [157, 19, 163].

5.1.2. Información geográfica voluntaria

Dentro de la generación de contenidos por parte del usuario están los sistemas de información geográfica de contribución voluntaria. Usuarios no profesionales generan contenido geoespacial utilizando para ello sistemas de mapeo disponibles en Internet, ofreciendo así posibilidades para que las agencias gubernamentales de todos los niveles mejoren sus bases de datos geoespaciales [165].

El término Información Geográfica Voluntaria (VGI, del término en inglés *Volunteered Geographic Information*) fue acuñado por Michael Goodchild para descri-

bir el uso de la Web para crear, reunir y difundir información geográfica provista voluntariamente por individuos [58]. Un ejemplo de este tipo de sistemas es OpenStreetMap.

Como se comenta en [85] este nuevo enfoque para la manipulación de información geográfica tiene dos ventajas principales:

- Permite tener mayor información disponible y más completa debido a las facilidades para la creación y edición del contenido por todos los usuarios. Esto es muy útil especialmente para el caso de regiones mal mapeadas, principalmente de los países en vías de desarrollo.
- Se puede obtener fácilmente los datos de forma gratuita, permitiendo así que los usuarios puedan crear mapas o utilizar los datos de una forma innovadora.

A pesar de los beneficios ya mencionados, esta forma de trabajar con datos geográficos también implica problemas relacionados con la calidad de los mismos [85, 6], problemas que se deben principalmente a dos causas:

- Los voluntarios que contribuyen a la creación del contenido carecen, generalmente, de la formación cartográfica necesaria para crear una buena representación de la información geográfica.
- La ausencia de un proceso que asegure la calidad de todas las operaciones de manipulación de datos.

Todo ello suele desembocar en entidades mal representadas o con una alta inexactitud. Por ejemplo, algunas de las entidades introducidas pueden asignarse a clases incorrectas debido a la interpretación individual de los datos o a un malentendido sobre las clases disponibles por parte de los voluntarios.

Muchas veces se argumenta que los problemas derivados de la calidad de los datos de los VGI representan una barrera para un uso más amplio por parte de las

agencias de mapeo oficiales. Sin embargo, investigaciones del *Center for Environmental and Geographic Information Services* han demostrado que algunos proyectos de mapeo participativo pueden producir datos que son tan precisos como los producidos por las agencias oficiales. Además, en algunos casos, los "ojos en el terreno" de los VGI tienen una ventaja sobre las pruebas de precisión más costosas de las agencias oficiales porque los contribuyentes tienen un conocimiento local único.

5.2 Fuentes de propósito general

En esta categoría se agrupan las dos fuentes de datos que son la columna vertebral de BITOUR, OpenStreetMap y Twitter. Se consideran de propósito general porque pueden utilizarse para obtener información y conocimiento en una amplia variedad de dominios, entre los los cuales se incluye el dominio de turismo que es el enfoque de esta plataforma.

5.2.1. OpenStreetMap

El proyecto OpenStreetMap (OSM) fue iniciado por Steve Coast en Inglaterra en el año 2004 como respuesta a los altos precios que cobraba la *Ordnance Survey*, la agencia cartográfica de Gran Bretaña, por su información geográfica. Desde entonces OSM se ha enfocado en fomentar el crecimiento, desarrollo y distribución de datos geo-espaciales de libre acceso así como garantizar libertad de uso y compartición de datos a cualquier usuario [124].

En la versión en español del articulo sobre OSM en Wikipedia[1] se listan las siguientes restricciones de acceso a datos geo-espaciales que tienen otras plataformas diferentes a OSM:

- En la mayoría de los países la información geográfica pública no es de libre uso.

[1] https://es.wikipedia.org/wiki/OpenStreetMap

- Las licencias de uso a veces restringen su utilización al tener el usuario un derecho limitado de aplicación de la cartografía. No se puede corregir errores, añadir nuevos datos o emplear esos mapas en integración en aplicaciones informáticas, publicaciones, etc. sin pagar por ello.

- En los últimos años han surgido iniciativas comerciales como *MapShare* de TomTom o *Map Maker* de Google orientadas a animar a los usuarios a completar, actualizar y corregir la cartografía de servicios privados. En la mayoría de estos casos los usuarios no tienen derecho alguno sobre la cartografía o sobre los datos añadidos o editados, pasando a ser sus contribuciones propiedad de dichas empresas (esto es, seguirá siendo cartografía propietaria y no libre)

OSM es un proyecto donde el contenido se crea de manera colaborativa por toda la comunidad, por lo que las personas que contribuyen a la creación del contenido constituyen el corazón del proyecto. Sin personas que recopilen y mantengan los datos, OSM pierde valor.

OSM tiene una naturaleza similar a Wikipidia y, de hecho, es considerada algunas veces como la Wikipedia de los datos geo-espaciales. En este sentido, OSM suele recibir las mismas críticas que Wikipedia en relación a la calidad de su contenido ya que cualquier persona puede crear o editar datos geo-espaciales. No obstante, y a pesar de las críticas a ambos proyectos (OSM y Wikipedia), estos se han consolidado como alternativas de primer orden a los sitios más tradicionales. En la figura 5.1 se puede apreciar como ha aumentado año tras año la cantidad de usuarios registrados en OSM que contribuyen al proyecto.

A diferencia de otros proyectos comerciales, un aspecto clave de OSM radica en que permite cualquier forma de reproducción de los datos y no es necesario solicitar permiso para ello. Los proyectos comerciales ofrecen derechos limitados a los usuarios sobre sus mapas (por ejemplo, si se quiere añadir una leyenda). En cambio, los datos de OSM se publican y se distribuyen bajo Licencia Abierta de

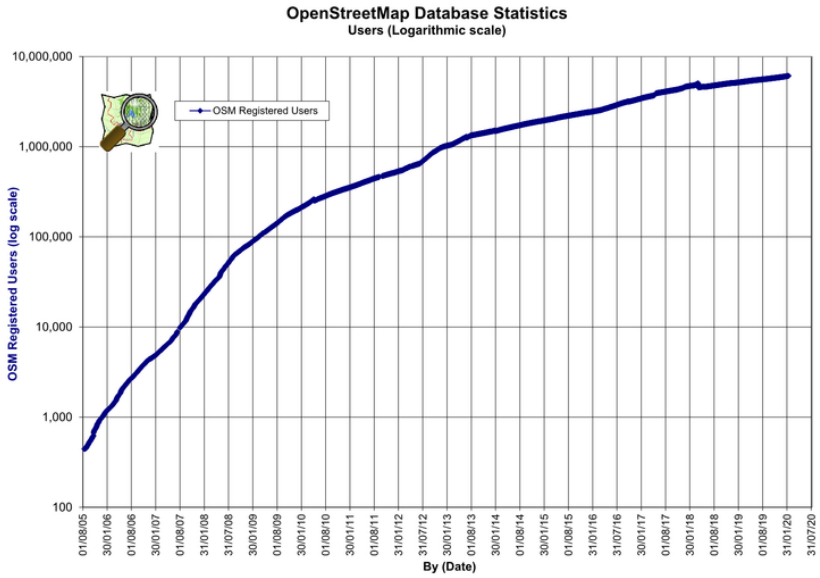


Figura 5.1: Usuarios registrados en OSM
(figura tomada de [124])

Bases de Datos (en inglés ODbL), lo que implica que cualquier versión modificada de los datos de OSM, si se usa públicamente, se deberá también compartir bajo una licencia ODbL.

El propósito central de OSM es recolectar datos geo-localizados y ponerlos a disposición de cualquier usuario en su formato básico (sin procesar). OSM ofrece también un conjunto de diferentes mapas en la web para acceder a los datos, siendo el más popular el sitio web del mapa del mundo [2]. Estas imágenes de mapas precalculados son de libre acceso y se pueden poner en cualquier sitio web utilizando pocas líneas de código *JavaScript*. La figura 5.2 muestra la visualización de la representación geo-espacial del Oceanografic ubicado en la ciudad de Valencia, España.

5.2.1.1. Formato de los datos

OSM utiliza una estructura de datos topológica. Los datos se almacenan en el datum WGS84 lat/lon (EPSG:4326) de proyección de Mercator. Los datos primitivos o elementos básicos de la cartografía de OSM son: **nodos** (también llamados puntos; algunas personas usan adicionalmente el término vértice); **ways** (lista or-

[2] `www.openstreetmap.org`

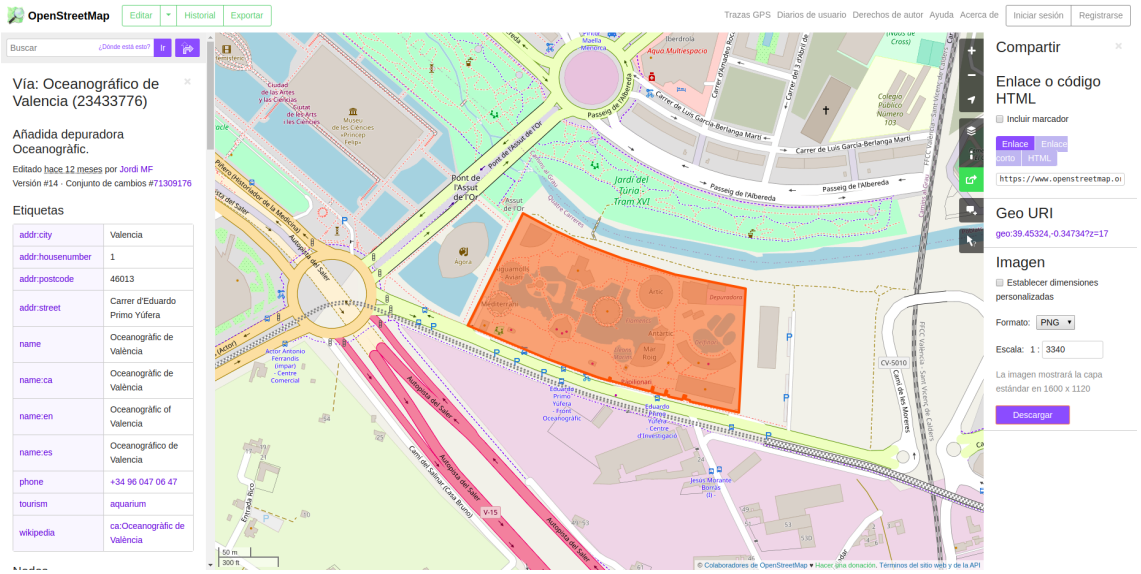


Figura 5.2: Oceanográfico de Valencia en OSM

denada de nodos que representa una línea o un polígono); y **relaciones**, las cuales se utilizan para modelar las asociaciones entre los diferentes elementos: nodos, *ways* y las propias relaciones.

Los nodos, *ways* y relaciones pueden ir asociados a una etiqueta (*tags*) que consta de una clave (*key*) y de un valor (*value*). Así, por ejemplo, la etiqueta `highway=trunk` define una vía como una carretera troncal. En la figura 5.3 se muestra la relación que existe entre los diferentes objetos (nodos, *ways* y relaciones) y sus etiquetas. En esta figura se puede apreciar que una relación consta de cero o más nodos, *ways* y relaciones, donde cada uno de ellos tiene un rol dentro de la relación y, además, pueden tener asociadas cero o más etiquetas. En la Tabla 5.1 se muestra un consolidado de las estadísticas de los objetos creados en OSM a la fecha de escritura de este documento.

Medida	Valor
Cantidad de usuarios	6.600.979
Cantidad de nodos	6.082.292.563
Cantidad de ways	671.357.311
Cantidad de relaciones	7.882.527

Tabla 5.1: Estadísticas OSM

Los **nodos** se utilizan para marcar Puntos de Interés (POIs) como, por ejemplo, la ubicación de una estación de gasolina, un museo, etc. Un nodo contiene la

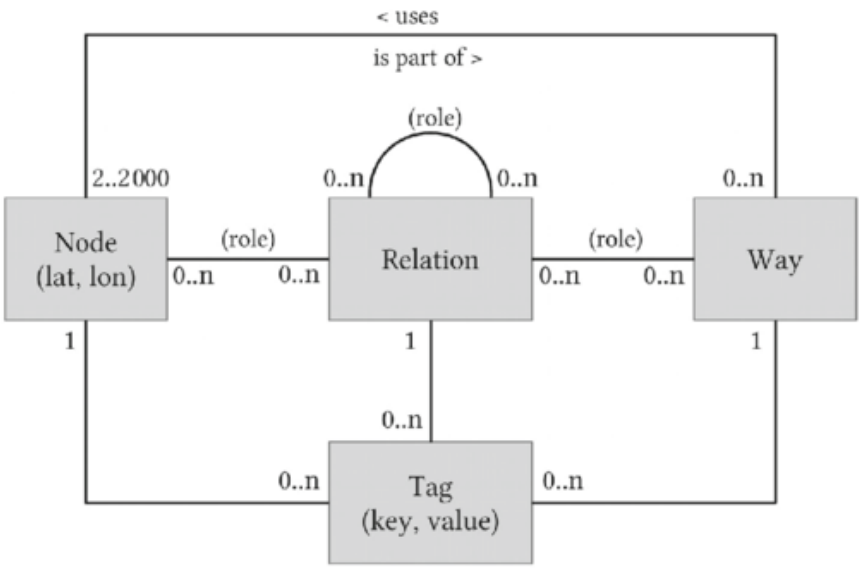


Figura 5.3: Modelo de datos de OSM
(figura tomada de [124])

Fragmento de código 5.1: Representation of a node in OSM

```
<node id="4320334552" visible="true" version="2" changeset="69685727"
timestamp="2019-04-29T08:22:16Z" user="Ornios" uid="733392"
lat="38.1415196" lon="-0.6462378">
...
    <tag k="information" v="office"/>
    <tag k="name" v="Oficina de Turisme"/>
    <tag k="name:en" v="Tourist Info"/>
    <tag k="name:es" v="Oficina de Turismo"/>
    <tag k="name:fr" v="Office du tourisme"/>
    <tag k="tourism" v="information"/>
...
</node>
```

siguiente información: identificador del elemento (ID), coordenadas geográficas (latitud y longitud), marca de tiempo de la última edición, nombre e identificador del usuario editor, versión del elemento, y cero o más etiquetas asociadas. El fragmento de código 5.1 muestra la representación de un nodo.

Los **ways** son el resultado de combinar nodos de distintas formas. Un *way* es una lista ordenada de nodos que representa una línea geométrica (un camino) o el perímetro de un polígono (empieza y finaliza en el mismo punto) como, por ejemplo, un lago. Un *way* consta de un identificador (ID), una marca de tiempo de la última edición, el nombre y el identificador de usuario que realizó la última edición, la versión del elemento, un identificador del conjunto de cambios que

se ha aplicado al elemento y una lista ordenada de al menos dos nodos y cero o más etiquetas. El fragmento de código 5.2 muestra un *way* que representa el Oceanografic ubicado en la ciudad de Valencia, España.

Fragmento de código 5.2: Representación de un *way* en OSM

```
1    <way id="23433776" visible="true" version="14" changeset="71309176"
2  timestamp="2019-06-16T20:06:35Z" user="Jordi MF" uid="8278438">
3
4      <nd ref="253802888"/>
5      <nd ref="253802889"/>
6      <nd ref="6551574249"/>
7      <nd ref="4638146471"/>
8      <nd ref="253802891"/>
9  ...
10     <tag k="addr:city" v="Valencia"/>
11     <tag k="addr:housenumber" v="1"/>
12     <tag k="addr:postcode" v="46013"/>
13     <tag k="addr:street" v="Carrer d'Eduardo Primo Yúfera"/>
14     <tag k="name" v="Oceanogràfic de València"/>
15     <tag k="name:ca" v="Oceanogràfic de València"/>
16     <tag k="name:en" v="Oceanogràfic of Valencia"/>
17     <tag k="name:es" v="Oceanográfico de Valencia"/>
18     <tag k="phone" v="+34 96 047 06 47"/>
19     <tag k="tourism" v="aquarium"/>
20     <tag k="wikipedia" v="ca:Oceanogràfic de València"/>
21 ...
22 </way>
```

Las **relaciones** son grupos de nodos, *ways* u otras relaciones a las que se pueden asignar determinadas propiedades comunes. Por ejemplo, una relación estaría formada por todos los *ways* que forman parte del edificio del Banco de España de la ciudad de Madrid (Fragmento 5.3). En este ejemplo se puede apreciar como la relación está conformada por siete *ways* donde cada uno tiene un rol: el *way* con identificador `194159908` representa un edificio interno como lo denota la etiqueta `role=''inner''`; mientras que el *way* con identificador `4487661` representa todo el perímetro del edificio tal y como indica la etiqueta `role=''outer''`. En la figura 5.4 se muestra la visualización de esta relación incluyendo todos los *ways*.

Los nodos, *ways* y relaciones puede estar asociados a un número de etiquetas. Las etiquetas están formadas por una clave y un valor, y ambos pueden ser una cadena de 255 caracteres y codificado en UTF-8. Las claves no pueden estar

Fragmento de código 5.3: Representación una relación en OSM

```
<relation id="2614014" visible="true" version="7" changeset="73818101"
timestamp="2019-08-28T01:38:16Z" user="Mapping4Fun" uid="5167321">

    <member type="way" ref="4487661" role="outer"/>
    <member type="way" ref="194159908" role="inner"/>
    <member type="way" ref="194159912" role="inner"/>
    <member type="way" ref="194159911" role="inner"/>
    <member type="way" ref="325705220" role="inner"/>
    <member type="way" ref="325705219" role="inner"/>
    <member type="way" ref="595928276" role="inner"/>
...
    <tag k="amenity" v="bank"/>
    <tag k="building" v="yes"/>
    <tag k="building:levels" v="5"/>
    <tag k="name" v="Banco de España"/>
    <tag k="office" v="government"/>
    <tag k="type" v="multipolygon"/>
    <tag k="wikidata" v="Q4889526"/>
    <tag k="wikipedia" v="es:Edificio del Banco de España (Madrid)"/>
...
</relation>
```

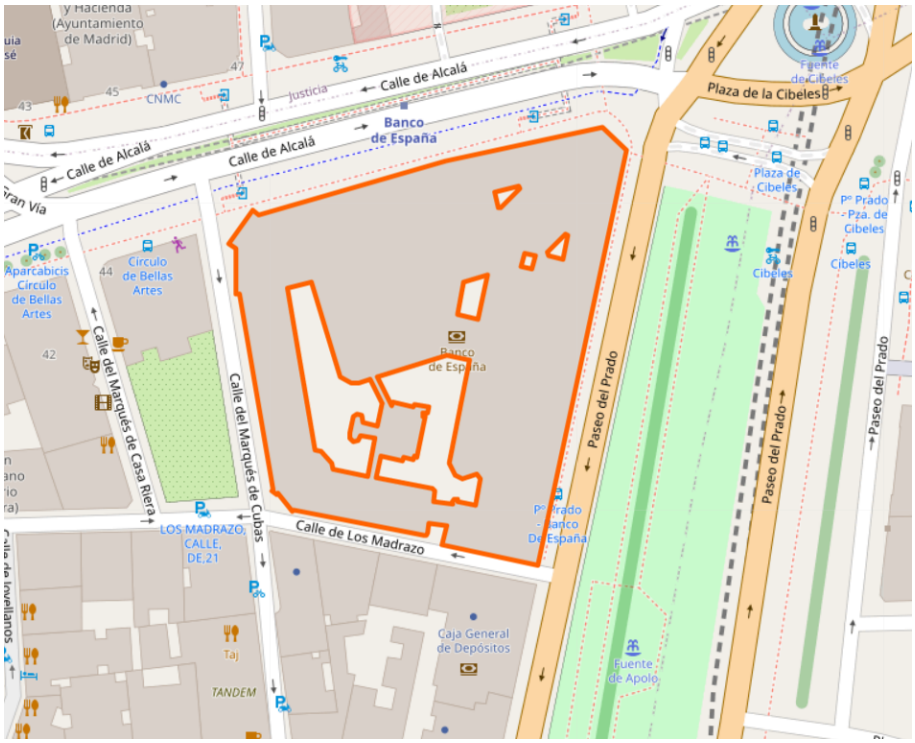


Figura 5.4: Banco de España en OSM

vacías, pero las valores sí, aunque esto último rara vez tiene sentido. Un objeto no puede tener múltiples etiquetas con la misma clave y las claves no distinguen entre mayúsculas y minúsculas.

5.2.1.2. Definición de las categorías

Para desarrollar el proceso de descarga de datos desde OSM, BITOUR utiliza el concepto de categorías. Estas son una forma de agrupar objetos de propósito similar, bajo un mismo nombre. Por ejemplo, agrupar bajo el nombre de categoría gastronomía los objetos que representan restaurantes y los objetos que representan bares, haciendo uso de las etiquetas de OSM. Para crear y editar las categorías BITOUR proporciona el formulario de la figura 5.5, el cual solicita la siguiente información:

- **Nombre de la categoría:** Se especifica el nombre de la categoría que permitirá agrupar a diferentes elementos de OSM. Es importante destacar que este nombre es independiente de los nombres que utiliza OSM. Por ejemplo, se puede definir el nombre de categoría Gastronomía. De acuerdo a esto, se puede definir "Gastronomía" como un nombre de categoría posible.
- **Etiquetas asociadas:** Los objetos de OSM son clasificados utilizando etiquetas que son la combinaciones de pares clave y valor. En esta funcionalidad, se aprovecha esta característica para permitir definir cuáles son las etiquetas ya existentes en OSM que permiten agrupar a los elementos de interés para el análisis en la categoría que se está definiendo. En la figura 5.5 se puede observar como a la categoría gastronomía son asociadas las etiquetas `amenity=bbq`, `amenity=restaurant` y `amenity=bar`.
- **Distancia:** Este dato no es relevante al momento de descargar la información de OSM pero tiene un uso posterior y se refiere a la distancia en metro que se utilizará para asignar otros objetos, como pueden ser los tuits realizados por un usuario a un objeto OSM definido bajo esta categoría. Se detallará el uso de este dato en el capítulo 6.

Con el uso de este formulario se puede crear, por ejemplo, la equivalencia entre las categorías de BITOUR y las etiquetas de OSM mostradas en la en la Tabla 5.2.

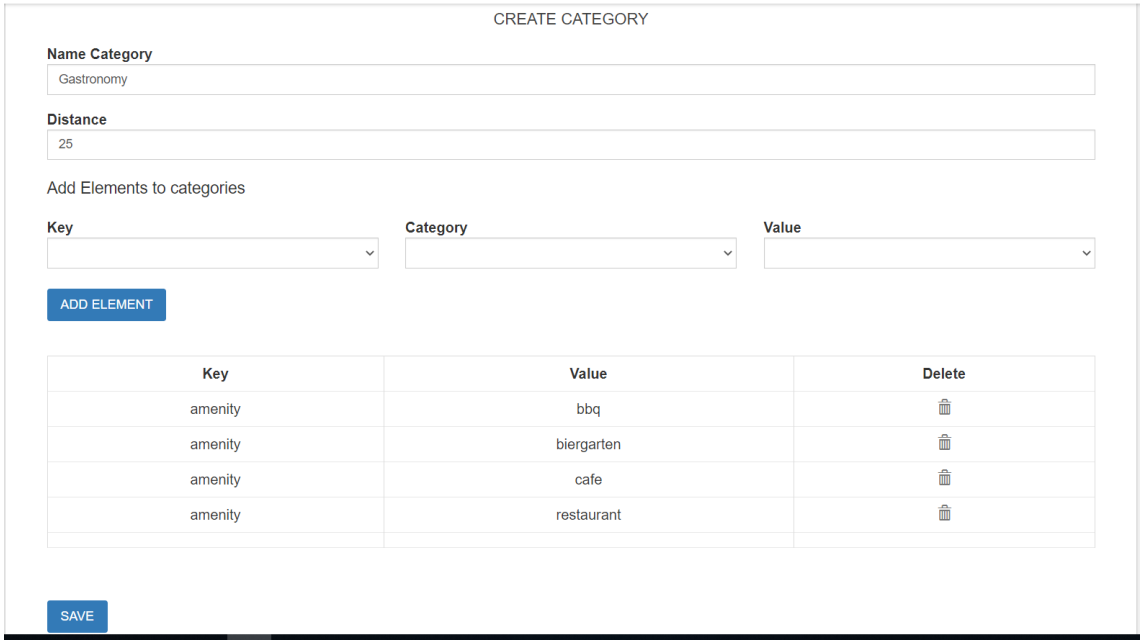


Figura 5.5: Definición de las categorías

5.2.1.3. Recuperación de información

La forma estándar para acceder automáticamente a los datos registrados en OSM es mediante el uso de la API Overpass (formalmente conocida como *OSM Server Side Scripting*). Las consultas a la API Overpass se pueden realizar tanto en formato Overpass XML o Overpass QL. Esta API es de solo lectura y permite extraer partes muy específicas de los datos de OSM. La API Overpass opera como se ve en la figura 5.6:

- Existe un servicio que habilita, a través de la Web, las datos de OSM. El servicio responde a los métodos públicos que la API Overpass implementa.

- El cliente envía una petición de datos en conformidad con los métodos públicos de la API Overpass utilizando el Protocolo de Transferencia de Hypertexto (HTTP, del inglés, *Hypertext Transfer Protocol*)

- El servicio evalúa la consulta y, si está en el formato apropiado, la ejecuta, recupera los datos desde el servidor y los devuelve al cliente en formato JSON.

- El cliente recibe los datos y es el responsable de su posterior procesamiento y/o almacenamiento.

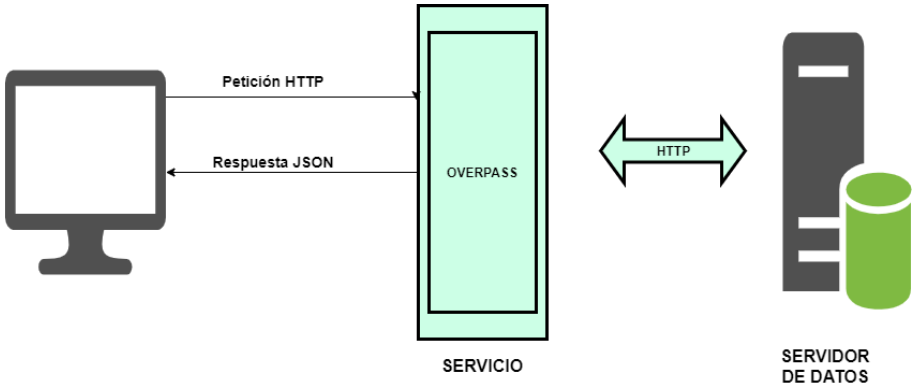


Figura 5.6: Interacción con la API Overpass

Los objetos que se dibujan en OSM están formados por una posición geométrica, una extensión y un tipo. El tipo se puede representar con nodos, *ways* o relaciones combinados con una o más etiquetas. Para obtener un objeto desde OSM (por ejemplo, un museo) se debe saber si dicho objeto fue dibujado como un nodo, un *way* o una relación y las etiquetas asociadas con él. La mayoría de los objetos de OSM se pueden describir usando solo un pequeño número de etiquetas. Por ejemplo, una ruta en OSM se puede describir utilizando la etiqueta `highway=footway` y el tipo *way*.

La información asociada a los objetos y la API Overpass se utilizan para recuperar datos desde OSM. El fragmento de código 5.4 muestra como se puede utilizar la API Overpass de OSM para obtener todas los *ways* ubicadas en Valencia y etiquetados como `highway=footway`.

La gama de etiquetas existente para clasificar los objetos OSM es amplia y no existe ningún control sobre las etiquetas que deben utilizarse en cada caso. El único control reside en la buena fe de los voluntarios. No obstante, OSM proporciona un listado exhaustivo de las etiquetas más utilizadas y que han sido aceptadas como un "estándar" no formal con el fin de que dichas etiquetas sean la primera

opción a utilizar en el momento de clasificar o recuperar un objeto. El listado se puede consultar en la página de OSM[3]

Fragmento de código 5.4: Consulta utilizando Overpass

```
1 [out:csv(::id,::type,"name")];
2     area[name="Valencia"];
3     way(area)[highway=footway];
4 out;
```

En este trabajo, el enfoque que se ha seguido para la recuperación de los datos consiste en utilizar la capa de abstracción que proporciona la plataforma BITOUR para agrupar las etiquetas de OSM bajo unas categorías más generales y que pueden ser de interés para el análisis turístico. De este modo, se puede, por ejemplo, agrupar las etiquetas `tourism=museum` y `amenity=art_center` bajo la categoría de museos. En la Tabla 5.2 se muestra las equivalencias que se han establecido entre las categorías turísticas utilizadas en este trabajo y las etiquetas de OSM.

El proceso de recuperación de datos de OSM se realiza para un destino en particular; dentro del destino se consulta cada una de las categorías turísticas creadas y para cada categoría se recupera las etiquetas asociadas.

Utilizando el caso concreto de la categoría 'Museos' (ver Tabla 5.2), podemos observar que dicha categoría está asociada a dos etiquetas. El proceso para recuperar la información es la siguiente:

- Se crea una consulta de datos utilizando la sintaxis de la API Overpass para cada una de las etiquetas que conforman la categoría. Se crea así una consulta para recuperar los objetos donde la clave `tourism` sea igual a `museum` y otra donde la clave `amenity` sea igual a `art_center`. En el fragmento de código 5.5 se muestra la consulta para el caso de la primera etiqueta.

- Por cada una de las consultas creadas se realiza un petición HTTP al servidor de datos de OSM, el cual se encarga de procesarla y recuperar los objetos

[3] `https://wiki.openstreetmap.org/wiki/Map_Features`

Tabla 5.2: Categorías de etiquetas de OSM

Categoría	Etiquetas OSM
Museos	(‘‘tourism’’, ‘‘museum’’); (‘‘amenity’’, ‘‘arts_centre’’)
Monumentos	(‘‘tourism’’, ‘‘attraction’’); (‘‘tourism’’, ‘‘viewpoint’’); (‘‘historic’’, ‘‘monument’’), (‘‘historic’’, ‘‘wayside_shrine’’), (‘‘historic’’, ‘‘memorial’’), (‘‘historic’’, ‘‘castle’’), (‘‘historic’’, ‘‘ruins’’), (‘‘historic’’, ‘‘archaelogical_site’’), (‘‘historic’’, ‘‘battlefield’’), (‘‘amenity’’, ‘‘grave_yard’’), (‘‘amenity’’, ‘‘crypt’’); (‘‘building’’,‘‘cathedral’’), (‘‘building’’,‘‘chapel’’), (‘‘building’’,‘‘church’’)
Sitios nocturnos	(‘‘amenity’’, ‘‘nightclub’’); (‘‘amenity’’, ‘‘pub’’), (‘‘amenity’’, ‘‘stripclub’’); (‘‘amenity’’, ‘‘bar’’)
Hoteles	(‘‘tourism’’, ‘‘hotel’’); (‘‘tourism’’, ‘‘hostel’’); (‘‘building’’,‘‘hotel’’)
Gastronomía	(‘‘amenity’’, ‘‘bbq’’), (‘‘amenity’’, ‘‘biergarten’’), (‘‘amenity’’, ‘‘cafe’’), (‘‘amenity’’, ‘‘restaurant’’)
Ocio	(‘‘tourism’’, ‘‘zoo’’); (‘‘tourism’’, ‘‘aquarium’’); (‘‘tourism’’, ‘‘theme_park’’); (‘‘amenity’’, ‘‘cinema’’); (‘‘amenity’’, ‘‘theatre’’); (‘‘leisure’’, ‘‘water_park’’); (‘‘leisure’’, ‘‘stadium’’); (‘‘leisure’’, ‘‘water_park’’); (‘‘leisure’’, ‘‘garden’’); (‘‘leisure’’, ‘‘park’’); (‘‘leisure’’, ‘‘playground’’), (‘‘leisure’’, ‘‘nature_reserve’’), (‘‘natural’’,‘‘beach’’); (‘‘natural’’,‘‘bay’’); (‘‘natural’’,‘‘cliff’’); (‘‘natural’’,‘‘coastline’’); (‘‘natural’’, ‘‘cave_entrance’’); (‘‘natural’’, ‘‘peak’’); (‘‘natural’’, ‘‘glacier’’); (‘‘natural’’, ‘‘volcano’’); (‘‘natural’’, ‘‘wood’’); (‘‘natural’’, ‘‘grassland’’); (‘‘natural’’, ‘‘tree’’)
Transporte	(‘‘aeroway’’, ‘‘aerodrome’’); (‘‘building’’,‘‘train_station’’)
Compras	(‘‘amenity’’, ‘‘marketplace’’); (‘‘shop’’, ‘‘mall’’)

de OSM que cumplen esa condición. Para el ejemplo de los museos en Valencia, España, se recuperan objetos como el museo fallero (ver fragmento de código 5.6). En este fragmento de código se observa que junto con el objeto viene información como la calle (`addr:street=‘‘Plaça Montolivet’’`).

- Los objetos recuperados se agrupan en un único paquete y se devuelve al cliente. Este último es el responsable de procesar los datos devueltos por cada una de las peticiones y guardarlos en el almacén de datos.

Fragmento de código 5.5: Consulta para recuperar los museos

```
[out:csv(::id,::type,"name")];
    area[name="Valencia"];
    way(area)[tourim=museum];
out;
```

Fragmento de código 5.6: Museo Fallero representado en OSM

```
<osm>
    <way id="444067498">
        <nd ref="4415706668"/>
        <nd ref="4415706669"/>
        <nd ref="4415706670"/>
        <nd ref="4415706671"/>
        <nd ref="4415706672"/>
        <nd ref="4415706673"/>
        <nd ref="4415706678"/>
        <nd ref="4415706674"/>
        <nd ref="1602559433"/>
        <nd ref="4415706675"/>
        <nd ref="4415706676"/>
        <nd ref="4415706668"/>
        <tag k="addr:city" v="Valencia"/>
        <tag k="addr:housenumber" v="4"/>
        <tag k="addr:postcode" v="46006"/>
        <tag k="addr:street" v="Plaça Montolivet"/>
        <tag k="building" v="yes"/>
        <tag k="building:levels" v="5"/>
        <tag k="name" v="Museo Fallero"/>
        <tag k="tourism" v="museum"/>
    </way>
</osm>

```

5.2.1.4. Uso de los datos de OSM

En poco más de una década, OSM se ha convertido en el principal ejemplo de VGI en Internet. OSM no es solo una base de datos geo-espaciales creados colectivamente sino que se ha desarrollado hasta el punto de convertirse en un vasto ecosistema de datos, sistemas de software y aplicaciones y herramientas que hacen uso de los datos contenidos en OSM [78, 159, 75].

El tipo de aplicaciones que hacen uso de los datos de OSM es muy diverso. A continuación se describe sola una pequeña muestra:

- Uno de los usos más frecuentes de OSM está orientado hacia la definición de rutas de tránsito (bicicletas, personas en sillas de ruedas y vehículos), permitiendo así a los usuarios de estos servicios obtener información oportuna y visual tanto de trayectos como de situación del tráfico [76, 95].

- Servicios de mapas para la búsqueda de lugares como casas, atracciones, restaurantes, etc. Estos servicios permiten responder a preguntas del tipo, ¿dónde esta el lugar X?, ¿cuál es la distancia entre los lugares A y B?, ¿cuál es el mejor trayecto entre dos puntos A y B? [14, 103, 33].

- Combinar la información de OSM con información de otras fuentes tanto de redes sociales [30] como de bases de datos enlazadas (*Linked Data*) [137, 121] para enriquecer la información de OSM.

- Utilización de los datos de OSM para la definición de políticas públicas [45] y creación de estadísticas oficiales [104].

5.2.2. Twitter

Twitter es un servicio de *microblogging* estadounidense, un sistema de comunicación consistente en un sistema de publicación de entradas de 280 caracteres máximo, y cuya información destaca por la simplicidad y la inmediatez. En esta red social los usuarios escriben e interactúan con mensajes conocidos como tuits.

Twitter se mantiene como una de las principales redes sociales donde los usuarios tuitean sobre cualquier tema dentro de los 280 caracteres permitidos y siguen a otros usuarios para recibir sus tuits. A diferencia de otras redes sociales en línea, tales como Facebook, la relación de seguimiento no requiere reciprocidad. Un usuario puede seguir a cualquier otro usuario sin necesidad que exista un seguimiento recíproco por parte de este último [88].

A nivel mundial, los datos estadísticos de Twitter durante el año 2019 se resumen del siguiente modo [145]:

- Hay más de 1.300 millones de cuentas creadas en Twitter.

- Hay 330 millones de usuarios activos mensuales en Twitter.
- Hay 152 millones de usuarios activos diarios en Twitter
- Hay 500 millones de tuits enviados por día
- Twitter es la plataforma más popular para consultar novedades informativas. El 79 % de las personas que buscan novedades utiliza esta red social.
- El 42 % de los usuarios de Twitter tiene presencia diaria en la red
- El usuario promedio pasa 3.39 minutos en la plataforma por sesión

5.2.2.1. Recuperación de información

Twitter es hoy en día el sitio de *microblogging* más ampliamente utilizado, lo que lo convierte en una valiosa fuente de información para analizar diferentes aspectos de un usuario como, por ejemplo, su ubicación, sus opiniones sobre un tema determinado, etc.

Para fomentar el uso de datos de Twitter, esta plataforma de información abierta proporciona puntos de acceso mediante los cuales se puede consultar y descargar sus datos. En este sentido, diversas iniciativas se han encaminado a aprovechar los datos disponibles en las redes sociales para realizar análisis en temas particulares como política (¿qué se piensa de un candidato?), creación de perfiles de usuario (¿a qué tipo de publicaciones se reacciona positivamente?), identificar la valoración que hace un usuario de una experiencia o relación con un negocio (los usuarios que han visitado un lugar dejan un comentario positivo o negativo del lugar), conocer la concentración de personas en un lugar (¿cuáles son los lugares más concurridos y con qué estacionalidad?) y las trayectorias que siguen los usuarios. Todas estas cuestiones pueden responderse a través de la información disponible de manera gratuita, legal y abierta de Twitter.

Twitter pone a disposición de los desarrolladores a nivel mundial una API que posibilita la recuperación de los tuits publicados y pertenecientes a personas

cuya cuenta es pública (la mayoría de las cuentas los son). Los puntos de acceso disponibles son los siguientes:

- Search API: permite buscar tuits publicados en los últimos 7 días.
- Ads API: crea una nueva campaña de anuncios de Twitter.
- Engagement API: permite obtener métricas de interacción de un tuit. Las solicitudes de API toman como entrada el identificador del tuit.
- Direct Message API: envía un mensaje directo.
- Account Activity API: recibe un mensaje directo a través de un *webhook*.
- Embed a Tweet: incrusta un tuit en un sitio web de elección.

Los tuits son el bloque de construcción atómico básico en Twitter. Todos los punto de acceso de Twitter que devuelven tuits proveen los datos usando la codificación de JSON. En esta codificación los tuits se representan como un objeto y los atributos del tuit se representan con la combinación clave/valor. De este modo, el objeto que representa un tuit puede tener atributos como `id=850006245121695744` y `text=''texto del tuit...''`. Adicionalmente, pueden existir objetos anidados dentro de los atributos del objeto tuit. Por ejemplo, se puede almacenar entro del atributo `user` de un tuit un nuevo objeto con sus propias atributos como `name=''Alex''` y `location=''Santa Marta''`. Es decir, un objeto de tuit contiene no solo información específica de los tuits, como el identificador, el texto o las coordenadas, sino también información del usuario que publicó el tuit, como el nombre, el país de origen, el idioma o la zona horaria. El fragmento de código 5.7 muestra un ejemplo de un extracto de un tuit.

Los atributos más relevantes que se pueden extraer de un tuit son:

- el `id` del tuit.
- `created_at` es el tiempo, en el formato `UTC`, cuando el tuit fue creado.

- `text`: el texto del mensaje en codificación `UTF-8`.

- `user`: el usuario que compartió el tuit – Este es un diccionario de datos anidado dentro de un objeto de tuit que incluye, entre otros atributos, el Id del usuario (`id`), la ubicación (`location`), el lenguaje (`lang`) especificado por el usuario y si este tiene habilitada la georeferenciación (`geo_enable`). En el fragmento de código se puede observar el diccionario de datos que representa esta información del usuario dentro del tuit.

- `coordinates`: representa la ubicación geográfica del tuit, la cual solo esta disponible si el usuario tiene activa la geolocaliación(`geo_enable`)

- `retweet_count`, representa el número de veces que el tuit ha sido retuiteado.

- `lang`: es el lenguaje que la plataforma asigna automáticamente a cada tuit. El valor asignado es uno de los valores agrupados por el `BCP`, `en` para inglés, `es-419` para el español hablado en Latino América y `undefined` en caso de que la detección automática no se pueda realizar.

Para la incorporación de datos de Twitter a la plataforma BITOUR se deben tener previamente descargados los tuits para un destino específico en un archivo de formato JSON. Estos datos se descargan utilizando la Search API que permite obtener los tuits realizados en una zona delimitada por una *bounding box*. Es así como se puede utilizar los puntos que definen los limites de un destino para obtener los tuits realizados dentro de estas zonas. Por ejemplo, para el caso de Valencia, la *bounding box* correspondiente a los valores `left=-0.7635`, `bottom=39.1701`, `right= -0.20050` y `top= 39.5883` se utiliza para obtener todos los tuits que se publican dentro de la zona definida por dichos puntos.

Una vez descargados los tuits se almacenan en un archivo de formato JSON que contiene un vector de objetos donde cada elemento corresponde a un objeto tuit como el listado en el fragmento de código 5.7. Para integrar los datos en la plataforma BITOUR se utiliza una rutina escrita en lenguaje PHP que toma el

Fragmento de código 5.7: Ejemplo de un extracto de un tuit

```
{
  "created_at": "Thu May 16 15:24:15 +0000 2019",
  "id": "850006245121695744",
  "text": "I am writing a paper...",
  "user": {
    "id": 2244994945,
    "name": "Alex",
    "screen_name": "TwitterDev",
    "location": "Santa Marta",
    "lang": "es",
    "geo_enable": "True"
    ...
  }
  "coordinates": [-3.51087576,39.46500176],
  "place": {
    "id": 2244994945,
    "place_type": "city",
    "name": "Valencia"
    ...
  }
  "lang": "en"
  ...
}
```

vector de objetos, recorre cada uno de los elementos (tuits), y por cada tuit extrae la información de interés.

La plataforma BITOUR tiene precargado los tuits de las ciudades de Valencia (España) y Berlín (Alemania). Estos datos se descargaron utilizando la Search API de Twitter durante el periodo comprendido entre febrero de 2015 y agosto de 2018.

5.2.2.2. Uso de los datos de Twitter

Los servicios de redes sociales han cambiando rápidamente la forma en que se crea, distribuye y comparte información [110]. Twitter se ha convertido en una valiosa fuente de información para distintos tipos de análisis, permitiendo extraer conocimiento [38] y responder a eventos casi en tiempo real [99]. Algunos de los usos más extendidos de los datos de Twitter en tareas de análisis son:

- Extraer el sentimiento expresado en el texto de los tuits mediante técnicas que van desde las más simples como bolsas de palabras hasta las más avan-

zadas como algoritmos de aprendizaje automático (máquinas de soporte vectorial, redes neuronales, etc.) [139].

- Utilización de la información geográfica que se puede extraer de los tuits para determinar los trayectos de los usuarios, sitios de concentración de personas y el tiempo de estadía en un lugar [96].

- Twitter también se ha utilizado en dominios específicos como el del turismo para conocer la imagen que tienen los turistas sobre un destino, identificación de turistas y residentes, etc. [4, 87, 24].

- Detección de *bots* en Twitter para identificar usuarios no humanos que escriben tuits de forma automática [7, 57].

5.3 Fuentes del dominio del turismo

En esta categoría se agrupan aquellas fuentes que proporcionan datos específicamente del dominio del turismo. Es decir, fuentes cuyo origen y utilización se destinan principalmente a proporcionar información de algún aspecto particular dentro del sector turístico. Las dos fuentes utilizadas son Tripadvisor y Airbnb.

5.3.1. Tripadvisor

Tripadvisor, la plataforma de viajes más grande del mundo, tiene un tráfico de más de 463 millones de visitantes únicos al mes. Viajeros de todo el mundo usan el sitio y la aplicación de Tripadvisor para consultar más de 859 millones de opiniones y comentarios sobre 8,6 millones de alojamientos, restaurantes, experiencias, aerolíneas y cruceros, ya sea que estén planificando un viaje o ya estén en uno. Los viajeros eligen Tripadvisor para comparar precios de hoteles, vuelos y cruceros, reservar *tours* y atracciones populares, así como también para reservar mesas en restaurantes. Tripadvisor se encuentra disponible en 49 países y en 28 idiomas [143, 160].

Tripadvisor fue fundada en Febrero del 2000 por Stephen Caufer. Los usuarios comparten experiencias y opiniones en Tripadvisor a nivel mundial y su uso se ha generalizado en diferentes ámbitos, siendo el turístico uno de los más beneficiados. De los diversos sitios web que proporcionan contenido generado por viajeros, Tripadvisor es la comunidad de contenido de viajes más grande del mundo [20]. Gracias al aumento de puntos wifi, los usuarios comparten sus experiencias en esta red social, permitiendo difundir cómo se sienten, sus gustos y dando opiniones sobre los servicios que reciben [109].

El portal ofrece diferentes servicios. algunos de los cuales se describen en la Tabla 5.3. Estos incluyen, entre otros, mapas para facilitar la ubicación de los usuarios y revisión de reseñas de los establecimientos y/o atracciones.

Tabla 5.3: Servicios ofrecidos por Tripadvisor

Características/Servicios	**Descripción**
Revisiones y calificaciones de viajes y/o destinos	Los viajeros pueden ver las opiniones y calificaciones generadas por otros viajeros y ver los perfiles de los revisores, la cantidad de revisores, etc. Este completo sistema de gestión de la reputación ayuda a los usuarios a determinar la utilidad de las revisiones y / o calificaciones. Lo que también permite a los usuarios analizar un lugar antes de estar allí.
Perfiles	Personalización: los usuarios pueden editar sus perfiles, para que puedan buscar y ver reseñas de viajes y sugerencias de acuerdo con sus preferencias y perfil de viaje.
Reseñas de un vistazo	Agregación de contenido: permite a los viajeros ver un resumen de las calificaciones de los viajeros, los tipos de viajeros y las últimas reseñas
Siendo tendencia ahora	Los viajeros pueden ver las últimas reseñas y contenidos agregados para un destino
Mapas	Mapas e información combinada: los mapas dinámicos visualizan diversa información relacionada con el viaje (por ejemplo, precio y disponibilidad del hotel) en un solo lugar.

5.3.1.1. Recuperación de información

Tripadvisor es un ejemplo de portal web que permite visualizar datos pero no proporciona medios para su descarga o acceso a través de una API gratuita. Por

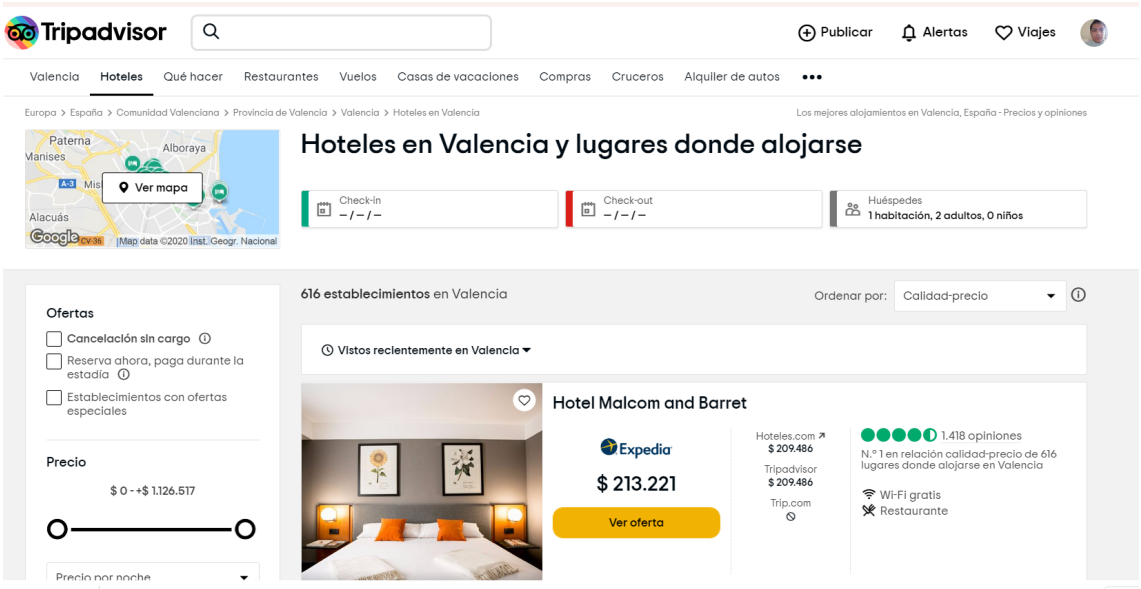


Figura 5.7: Sitio web de Tripadvisor

tanto, el acceso a la información de Tripadvisor no se puede automatizar de una forma sencilla. En la figura 5.7 se puede observar el resultado de buscar los hoteles ubicados en la ciudad de Valencia, España, allí se muestra para el hotel "Malcom and Barret" información como la valoración, precio, cantidad de revisiones, etc.

Para solventar la dificultad de acceso a los datos de Tripadvisor se ha utilizado la técnica conocida como *web scraping*, la cual consiste en la transformación de datos web semi estructurados (como el formato HTML) en datos estructurados que pueden ser almacenados y analizados en una base de datos central, en una hoja de cálculo o en algún otro destino (ver figura 5.8).

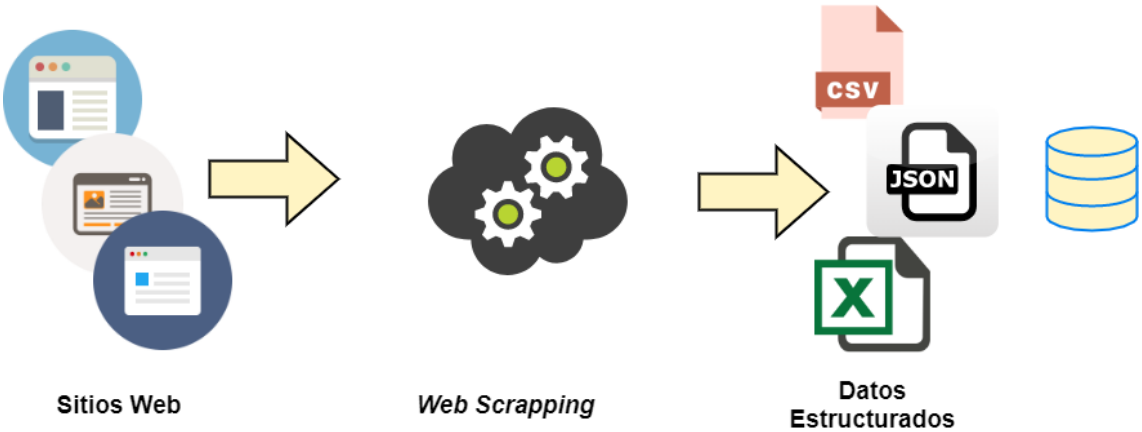


Figura 5.8: *Web scrapping*

La técnica *web scrapping* aprovecha que los sitios web se escriben usando el lenguaje HTML, lo que significa que toda página web es un documento estructurado con secciones claramente definidas. Es importante aclarar que el HTML da estructura al contenido de la página más no a los datos que en ella se muestran. La estructuración del documento se consigue a través de un sistema de marcas o etiquetas; por ejemplo, para el título se utiliza la etiqueta `Title`, para un párrafo la etiqueta `P`, entre otras etiquetas. Además, cada una de estas etiquetas puede tener un nombre (`name`) un identificador (`id`) y una clase (`class`) que permite cualificarlas y acceder a ellas de una manera más directa. En el extracto de código 5.8 se presenta un ejemplo de la estructura de una página web básica donde se puede observar que:

- La pagina web se constituye a partir de etiquetas como `html`, `body` y `title`. Cada una de estas etiquetas tiene un significado para un navegador web, el cual es el responsable de su visualización.
- Las etiquetas pueden estar anidadas, es decir, algunas etiquetas pueden contener una o más etiquetas de su mismo tipo o de otro tipo.
- El contenido de algunas etiquetas, como `title`, es texto que se muestra directamente en la ventana del navegador. Por ejemplo, la etiqueta del título contiene el texto “Soy el título”.
- Las diferentes etiquetas conforman una representación de la relación existente entre los diferentes objetos de la página que se denomina Modelo de Objetos del Documento (DOM, del término *Document Object Model*). El DOM muestra cómo están anidadas las etiquetas y en qué nivel se encuentra cada una de ellas. La figura 5.9 muestra el DOM del extracto de código 5.8, donde se puede observar que el elemento raíz es `html`, que a su vez tiene dos hijos (`head` y `body`); y que el `body` también tiene dos hijos, un párrafo y una cabecera de primer nivel, conteniendo esta última el texto “esto es una cabecera nivel uno”.

Fragmento de código 5.8: Representación de HTML

```
<html>
<head>
  <link rel="stylesheet" href="estilos.css">
  <title>Soy el título</title>
</head>
<body>

    <h1>Esto es una cabecera nivel uno</h1>
    <p id="precio">Esto es un párrafo</p>

</body>
</html>
```

Es importante hacer explícito que debido a que los datos extraídos desde esta fuente son únicamente los hoteles (coordenadas, nombre, precio y valoración) que aparecen listados en una misma página (con paginación), el número de peticiones para descargar los hoteles de un destino, generalmente, no dispara las alertas de lo que Tripadvisor considera un robot. No obstante, para cuando esto ocurre, la estrategia que se maneja es pausar el proceso de descarga, esperar 24 horas (el tiempo de bloqueo) y reiniciar la descarga con los hoteles restantes.

Es resumen, para acceder a la información de Tripadvisor se debe acceder al código HTML de la página, conocer la estructura de los elementos y utilizar el DOM para navegar por las diferentes partes de la página y extraer la información necesaria. De este modo, para extraer el número de opiniones sobre el "Hotel Malcom and Barret" (ver figura 5.10) se debe buscar el elemento HTML que tiene como clase `review_count` para luego obtener su contenido `<a class=review_count>1.418 opiniones </a>`. Este procedimiento debe repetirse para cada dato que se desee extraer.

5.3.1.2. Uso de los datos de Tripadvisor

Debido a la amplia popularidad de Tripadvisor entre los turistas y a la gran cantidad de información que maneja, sus datos han sido y son de interés para diferentes análisis y aplicaciones [32, 22]. A continuación, se describen algunos ejemplos de estas aplicaciones:

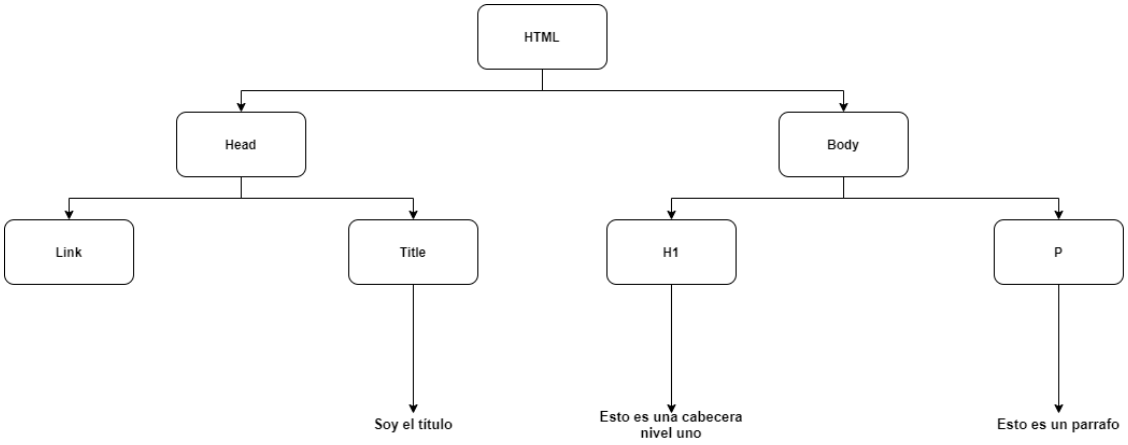


Figura 5.9: Modelo de objetos del documento

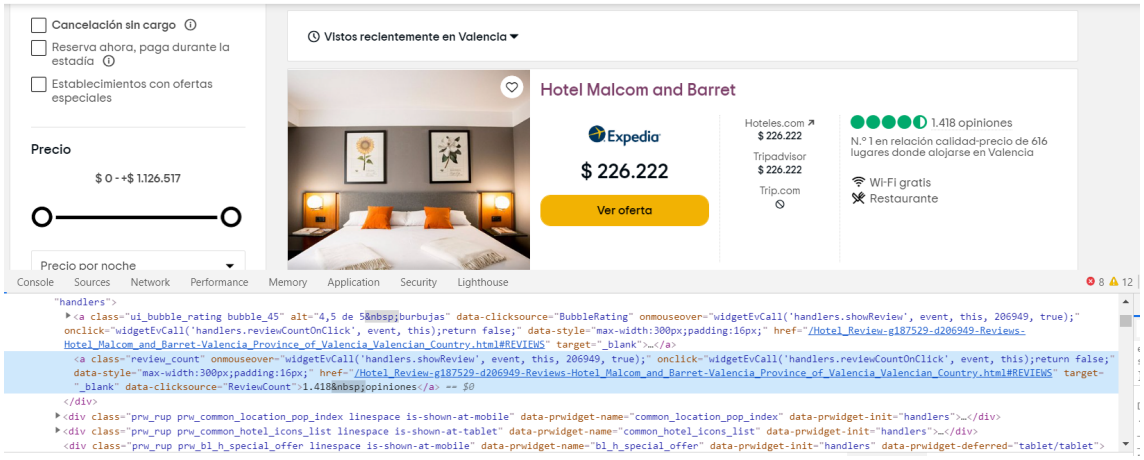


Figura 5.10: Código HTML de Tripadvisor

- Creación de perfiles de turistas que visitan los establecimientos que figuran en Tripadvisor y dejan un comentario en el portal [21, 141, 92].

- Uso de los comentarios hechos en Tripadvisor para crear sistemas de recomendación con el objetivo de sugerir sitios a visitar para futuros turistas [63, 18, 140].

- Descubrir la imagen pública que tienen determinados establecimientos (por ejemplo, hoteles [120, 90]). Es importante conocer el efecto de las opiniones de los turistas en Tripadvisor sobre lugares de alojamiento, restauración, etc. y la influencia de estas opiniones sobre las visitas que dichos establecimientos reciben. [158].

5.3.2. Airbnb

Este proyecto, cuyo nombre original era *AirBed and Breakfast*, fue fundado por tres egresados universitarios a mediados de 2008. En sus inicios el proyecto era un sitio web que ofrecía solo espacios compartidos y habitaciones privadas (pero no casas enteras) y su propósito era proporcionar alojamiento durante eventos importantes. El servicio Airbnb ha evolucionado rápidamente hasta el punto de convertirse en el portal de alquiler de alojamiento más inclusivo existente a día de hoy [62]. Aunque inicialmente su popularidad fue limitada, a mediados de 2010 la trayectoria de crecimiento de la compañía comenzó a aumentar bruscamente, y ha continuado por el mismo camino desde entonces. Según la pagina web del proyecto[4], a fecha de hoy tiene más de 7 millones de lugares que prestan servicio de alojamiento y Airbnb está en más de 100 mil ciudades y 200 países.

Airbnb se describe a sí mismo como "un mercado comunitario de confianza para que las personas enumeren, descubran y reserven alojamientos únicos en todo el mundo". Es esencialmente una plataforma en línea a través de la cual las personas interesadas en alquilar sus espacios para turistas, y los turistas interesados en alquilar un espacio, pueden encontrarse y satisfacer sus necesidades. Los alojamientos Airbnb generalmente involucran una casa completa (como un apartamento o casa) o una habitación privada en una residencia donde el anfitrión también está presente. Un porcentaje muy pequeño de las listas de Airbnb son habitaciones compartidas (por ejemplo, un huésped puede dormir en un espacio en la sala de estar) o alojamientos exóticos como iglús y casas en los árboles. El diverso inventario de Airbnb también varía desde alojamientos muy modestos hasta extremadamente lujosos.

El surgimiento de Airbnb es, sin duda, uno de los desarrollos recientes más significativos y transformadores dentro del sector turístico mundial. Aunque Airbnb lleva en funcionamiento aproximadamente solo unos 10 años, ha conseguido re-

[4] `https://news.airbnb.com/fast-facts/`

volucionar la antigua práctica de alojamiento. Esto es visto de manera positiva y negativa por algunas entidades [61]:

- La compañía ha desatado una oportuna innovación que ha crecido más rápido de lo que prácticamente nadie habría esperado, transformando a innumerables personas en micro empresarios hoteleros.
- El éxito y la expansión de Airbnb se ha convertido, a su vez, en un problema político por el modo en que afecta a alojamientos turísticos tradicionales.

Los principales servicios ofrecidos por Airbnb y que marcan la diferencia con respecto a otros servicios de alojamiento son:

- A diferencia de las plataformas de emparejamiento puro como Craigslist o plataformas de distribución como Expedia, Airbnb participa en numerosos aspectos de las transacciones que facilita. Airbnb procesa los pagos de los huéspedes a los anfitriones y gana dinero cobrando una "tarifa de servicio" (porcentaje de comisión) de ambas partes.
- Airbnb también alienta a huéspedes y anfitriones a que valoren públicamente a la otra parte, lo que ayuda a fomentar la confianza subyacente necesaria para que un servicio de este tipo prospere.
- Airbnb promueve aún más la confianza y la seguridad al ofrecer varias medidas de verificación de identidad, protección gratuita contra daños a la propiedad ("Garantía del Anfitrión"), seguro de responsabilidad civil gratuito ("Seguro de Protección del Anfitrión") y una "Política de reembolso del huésped".
- Airbnb también se ha extendido más allá del alojamiento turístico, y ahora procesa adicionalmente reservas de restaurantes y ofrece "experiencias", que incluyen recorridos u otras excursiones dirigidas por guías locales.

- Además, Airbnb se ha asociado con varias compañías de gestión de viajes para facilitar los viajes corporativos. Recientemente incluso se ha asociado con un desarrollador de bienes raíces para construir complejos de apartamentos diseñados para el alquiler de Airbnb.

5.3.2.1. Recuperación de información

Airbnb permite el acceso a información sobre una gran cantidad de alojamientos en diversos destinos. En la figura 5.11 se puede observar una muestra del listado de sitios que la plataforma despliega al buscar información sobre alojamiento en la ciudad de París, Francia. Dentro de la información que despliega Airbnb para los alojamientos mostrados se encuentra el precio, la ubicación, la valoración de otros clientes. La figura 5.12 muestra los detalles de un alojamiento seleccionado como los comentarios que ha recibido, información sobre los servicios y número de personas que lo han valorado, entre otros aspectos.

Al igual que Tripadvisor, Airbnb no proporciona una manera sencilla de acceder a los datos a través de una API. La recuperación de datos de Airbnb tiene que hacerse a través del código HTML de la página aunque es importante mencionar que el proyecto Inside Airbnb[5] proporciona los datos de algunas ciudades, principalmente europeas, de renombre mundial como es el caso de París, Roma, Madrid, Valencia, entre otras muchas más.

En este trabajo de tesis doctoral se ha utilizado los datos proporcionados por el proyecto Inside Airbnb, los cuales están en un formato tabular, particularmente en formato de archivo CSV. El procesamiento es simple, se carga el archivo en memoria y se procesa línea a línea hasta extraer los datos necesarios. El proceso es el que se describe a continuación:

- Se recorre cada uno de los establecimientos registrados en Airbnb del archivo CSV cargado correspondiente a la ciudad o área de destino que se quiere analizar.

[5] `http://insideairbnb.com/get-the-data.html`

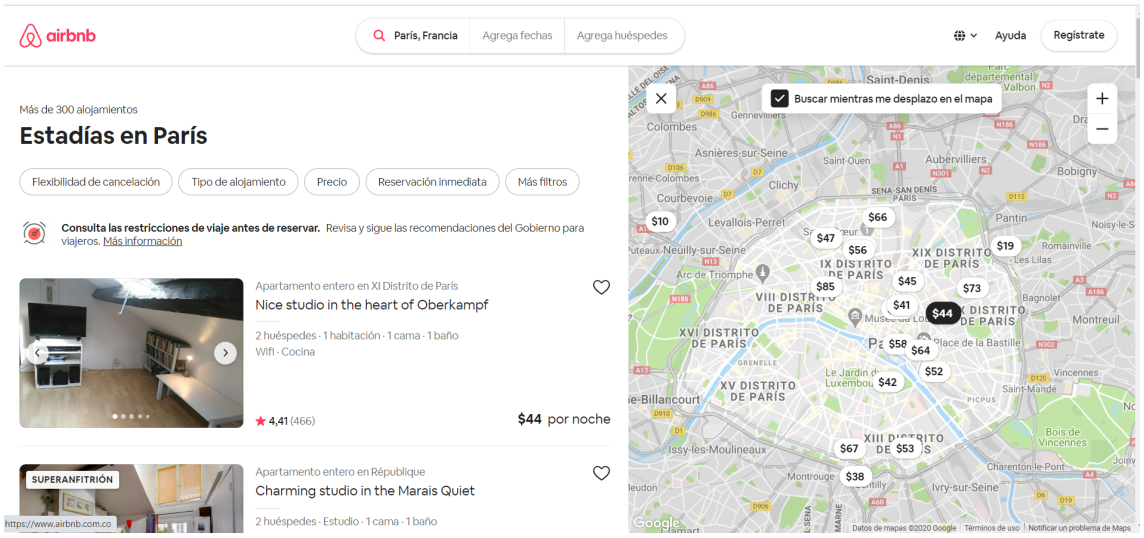


Figura 5.11: Sitio web de Airbnb

Figura 5.12: Detalle de un alojamiento en Airbnb

- Por cada fila se extraen los valores pertenecientes al nombre, el precio de una noche en el establecimiento, la puntuación media de los usuarios y las coordenadas geográficas del lugar.

5.3.2.2. Uso de los datos de Airbnb

Airbnb se ha convertido en uno de los modelos de negocio más disruptivos en el ámbito del sector turístico a escala internacional y constituye uno de los ejemplos de éxito de los llamados negocios peer-to-peer [8, 151]. Al igual que Tripadvisor esta fuente proporciona datos que son valiosos para realizar diferentes tipos de análisis en el dominio del turismo. A continuación se listan algunos usos que se han hecho de los datos que Airbnb proporciona:

- Análisis comparativo de los niveles de satisfacción de los turistas alojados en establecimientos de Airbnb con respecto a los que han elegido otro tipo de establecimientos de carácter más convencional [8, 60].
- Utilización de técnicas de minería de texto y minería de opiniones para identificar los atributos que influyen en la experiencias de los turistas en un alojamiento y/o destino [31, 62].

5.4 Justificación de la selección

Las razones que nos han llevado a seleccionar las cuatro fuentes de datos expuestas en este capítulo son las siguientes:

- OSM: es el proyecto líder y más completo de información geográfica y gratuita a nivel mundial. Además, proporciona mecanismos de acceso automático a la información espacial de un lugar de interés a través de APIs.
- Twitter: Es la red social de *microbloging* más popular y utilizada a nivel mundial, convirtiéndose así en una valiosa fuente de información sobre opiniones personales. Una ventaja que proporciona Twitter es que la información de usuarios y sus opiniones está geo-referenciada.

- Tripadvisor: Existen varias plataformas y agencias virtuales de viaje que proporcionan la misma información que Tripadvisor (precio, valoración, nombre y ubicación de alojamientos), pero la selección de Tripadvisor se hizo porque esta plataforma tiene la mayor cantidad de contenido creado por usuarios en cuanto a revisiones y valoraciones de los establecimientos.

- Airbnb: Es el proyecto líder a nivel mundial en cuanto a la prestación de servicio de intermediación entre anfitriones y viajeros para alojamiento informal, generalmente con fines turísticos. Proporciona datos importantes de los establecimientos tales como el precio, valoraciones personales y ubicación de los lugares. La utilización de Airbnb permite completar la información de Tripadvisor, cuyo segmento de información es el alojamiento formal (hoteles y hostales).

5.5 Resumen

En este capítulo se describieron las fuentes de datos utilizadas en BITOUR. Dichas fuentes son clasificadas en dos categorías: fuentes de propósito general como OSM y Twitter y fuentes de propósito específico como Tripadvisor y Airbnb. De cada una de estas fuentes se enunció el tipo de aporte en materia de información a BITOUR: OSM proporciona la información espacial de los lugares dentro de un destino; Twitter la opinión de los usuarios sobre el destino y Tripadvisor y Airbnb información sobre los sitios que prestan servicios de alojamiento bien sean hoteles o apartamentos turísticos.

Posteriormente, de cada una de las fuentes se describió cómo están organizados sus datos y el método para acceder a ellos. Así, en el caso de OSM, se utilizó la API *Overpass* junto con la capa de abstracción de categorías que proporciona BITOUR para cargar los datos de interés; con Twitter se utilizó la API de desarrolladores para precargar los datos de Valencia (España) y Berlín (Alemania); en el caso de Tripadvisor se utilizó el *Webscrapping* para descargar la información de los

hoteles; y con Airbnb, se realizó una lectura de datos estructurados en formato CSV que proporciona el sitio web *Airbnb inside*.

CAPÍTULO 6

Procesamiento y visualización de datos

En este capítulo se describen aquellos procedimientos de la plataforma BITOUR que necesitan ser profundizados con el objeto de proporcionar claridad sobre cómo opera la plataforma, qué se necesita para poder utilizar a los turistas y los tuits como unidades de análisis y cómo se pueden explotar estos datos. Para ello, el capítulo se divide en dos secciones. En la sección 6.1 se explican detalles como qué criterios se utilizan para asignar los tuits a los lugares almacenados en la plataforma y cómo se cataloga a un usuario como un *bot* o como turista. En la sección 6.2 se presenta cómo operan las diferentes componentes de visualización que la plataforma pone a disposición de los analistas.

6.1 Detalles de la capa de procesamiento

El propósito de esta sección es ofrecer mayor detalle con respecto a tres tareas que son claves para el correcto funcionamiento de BITOUR: inicialmente la sección 6.1.1 muestra la manera como los tuits son asignados a los lugares cargados previamente en un destino; luego la sección 6.1.2 describe qué criterios se utilizan para identificar aquellos usuarios que no son humanos (*bots*); la sección 6.1.3 muestra la técnica aplicada para distinguir entre usuarios que son turistas y aquellos que son residentes del destino.

6.1.1. Procedimiento de asignación de los tuits

Este procedimiento constituye una piedra angular para el funcionamiento de BITOUR debido a que sirve de soporte tanto para posteriores tareas de visualización como de datos de entrada para otras tareas, como la identificación de turistas. El propósito de este procedimiento es asignar un tuit a un lugar con el objetivo de saber desde qué lugar se realizó el tuit. La razón por la que es importante esta asignación radica en que la categoría a la que están asignados los lugares es usada con posterioridad tanto para la identificación de los turistas como para la visualización de los datos.

Para ello, el procedimiento se basa en las siguientes premisas:

- **Todos los lugares están agrupados en categorías que denotan el tipo de actividad que se puede realizar en ellos.** Es así como unos lugares pueden ser categorizados como museos, otros como monumentos, etc.
- **Se considera que un tuit se realiza desde un lugar en particular si la distancia entre la ubicación del tuit y la ubicación del lugar es inferior a un valor, en metros, establecido previamente para cada categoría.** Por ejemplo, si hemos definido que la distancia máxima permitida para considerar que un tuit se realizó desde un hotel es de 35 metros, todo tuit a una distancia menor o igual a 35 metros con respecto a un lugar previamente categorizado como hotel puede ser asignado al lugar.
- **Cada tuit solamente puede ser asignado a un lugar.** Sin embargo, puede ocurrir que, dado un tuit determinado, éste pueda ser asignado a más de un lugar porque cumplen la condición de distancia máxima. Por ello, se define una lista de prioridades por categoría de manera que el tuit se asignará al lugar de máxima prioridad que cumpla la condición de distancia máxima.

Por consiguiente, el primer paso es saber por cada categoría definida cuál es la distancia máxima permitida para considerar que un tuit pertenece a un lugar de

esta categoría y el orden de prioridad que existe entre las diferentes categorías. A modo de ejemplo, la tabla 6.1 muestra la definición de ocho categorías con la distancia máxima permitida y su prioridad (la máxima prioridad es 1). Por ejemplo, la distancia máxima permitida con los museos es de 25 metros, siendo esta categoría la de máxima prioridad.

Tabla 6.1: Ejemplo de prioridades y distancias usadas por categorías

Categoría	Distancia	Prioridad	Prioridad	Distancia	Prioridad
Museos	25 metros	1	**Gastronomía**	25 metros	5
Monumentos	50 metros	2	**Ocio**	25 metros	6
Ocio Nocturno	25 metros	3	**Transporte**	15 metros	7
Hoteles	35 metros	4	**Compras**	15 metros	8

Los datos de entrada al procedimiento de asignación de tuits serán, por tanto:

- Tuits con ubicación geográfica pendientes de asignación.
- Lugares con ubicación geográfica cargados en la plataforma y clasificados en categorías.
- Categorías con una distancia y prioridad preestablecida.

Para realizar esta tarea, BITOUR dispone de dos formularios que se describen a continuación. La figura 6.1 muestra el formulario que lista las categorías definidas para el destino y da la opción para crear nuevas categorías y editar las existentes; la figura 6.2 muestra la información que debe ser ingresada para crear una nueva categoría: el nombre, la distancia máxima permitida y los objetos asociados a ella. Ambos formularios son utilizados por el administrador.

Con esta información, el procedimiento de asignación funciona de la siguiente forma:

- Por cada tuit se toma su ubicación y se calcula la distancia que existe entre cada tuit y los lugares guardados en la plataforma y se verifica que cumplan con la condición de estar dentro de la distancia máxima permitida, conservando únicamente aquellos lugares que satisfacen esta condición.

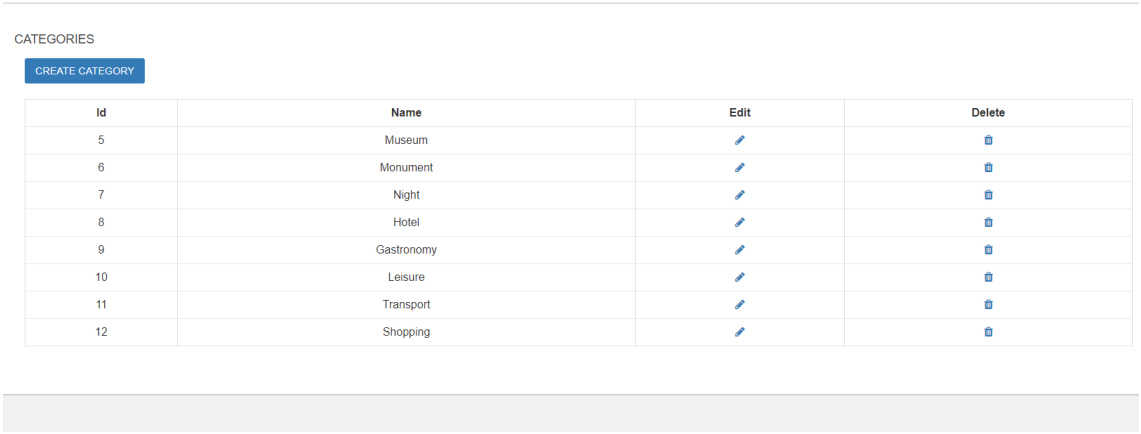


Figura 6.1: Listado de categorías creadas

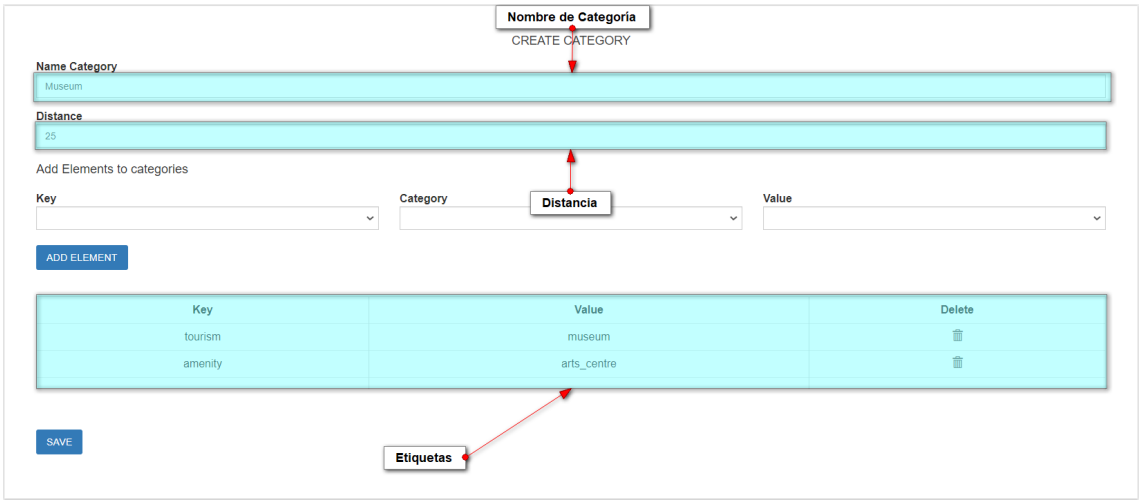


Figura 6.2: Configuración de cada categorías

- De los lugares que satisfacen el criterio de distancia, se selecciona por cada categoría el lugar más cercano al tuit. Posteriormente, de todos los lugares restantes, se selecciona el que pertenece a la categoría de mayor prioridad.
- Si tras ejecutar los dos pasos anteriores al menos un lugar satisface ambos criterios, el tuit se asigna a este lugar. De lo contrario, el tuit queda sin asignar.

A modo de ejemplo se puede considerar el tuit de identificador 1020: tras calcular la distancia entre su ubicación y la de los lugares almacenados en la plataforma, se obtienen los valores reflejados en la tabla 6.3. En esta tabla se puede apreciar cómo hay tres lugares que satisfacen el criterio de distancia, uno de la categoría Monumentos, uno de la categoría Ocio y otro de la categoría Gastronomía.

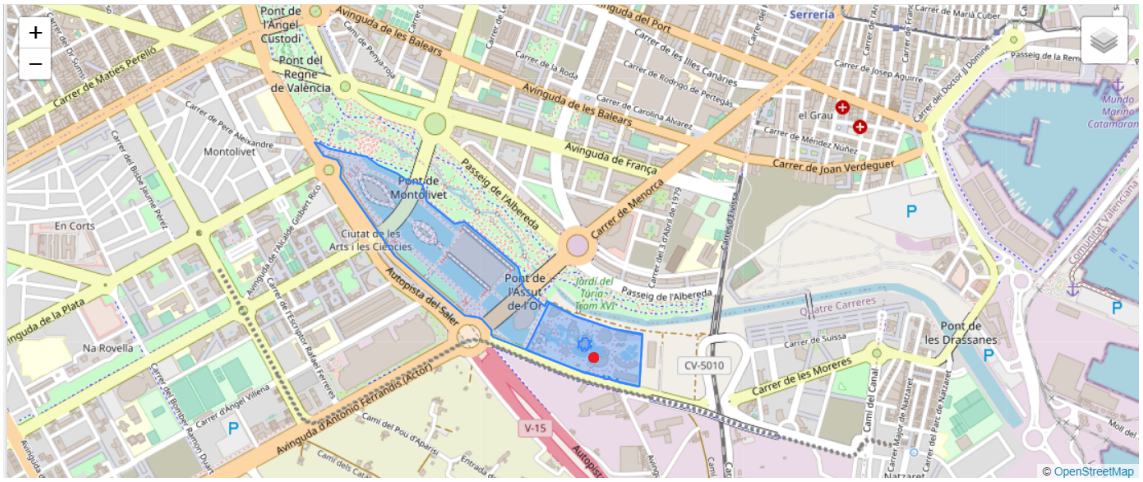


Figura 6.3: Visualización geográfica del tuit

Por consiguiente, siguiendo el criterio de prioridad, el tuit es asignado al lugar categorizado como Monumento, ya que su prioridad es 2 frente a las prioridades 5 y 6 de Gastronomía y Ocio.

Tabla 6.2: Ejemplo de prioridades y distancias usadas por categorías

Id tuit	Nombre de lugar	Distancia	Categoría
1020	Restaurante Submarino	5,89 metros	**Gastronomía**
1020	Ciudad de las artes y de las ciencias	0 metros	**Monumento**
1020	Oceanográfico	0 metros	**Ocio**

Es preciso aclarar que esta forma de realizar las asignaciones puede desencadenar en tuits ubicados en lugares que están muy próximos, por lo será importante en trabajos futuros incorporar otros elementos que nos ayuden a tener mayor precisión. Por ejemplo, realizar la detección de las entidades a las cuales se hace mención en el texto del tuit y considerar la hora del tuit para estimar la probabilidad de que un tuit sea hecho desde un lugar u otro. A pesar de estas limitaciones, se poseen ventajas como el rendimiento computacional.

6.1.2. Procedimiento para la detección de los bots

Un paso importante para reducir el ruido al que pueden inducir usuarios no reales es el de detectar estos usuarios y excluirlos del análisis de manera que se reduzca el riesgo de obtener estadísticas poco reales. BITOUR permite ejecutar un procedimiento que consiste en identificar aquellos usuarios que no correspon-

den a una persona que realiza tuits, sino a una máquina que lo hace de manera programada.

Para la detección de aquellos usuarios que puede ser considerados como *bots* se sigue el siguiente proceso:

- Por cada usuario, se toma su `id` y se extraen todos los tuits que el usuario ha realizado.
- Por cada uno de los tuits de un usuario se calcula la distancia que existe entre ese tuit y los demás tuits, hasta obtener la distancia entre cada par de tuits realizados por el usuario.
- Una vez se tiene la distancia entre todos los tuits que el usuario ha realizado, se utiliza como criterio para decidir si el usuario es un *bot* el hecho de que tenga al menos 10 tuits realizados desde una distancia menor a los 20 metros.

A modo de ejemplo podemos considerar el usuario identificado con el `id` `1011001099`. Este usuario tiene registrados 15 tuits en un periodo de 30 días desde la ciudad de Valencia, España. El siguiente paso es calcular la distancia que hay entre cada uno de sus tuits. En la figura 6.4 se muestra cómo están distribuidos los tuits (puntos azules) del usuario. En la figura da la apariencia que no están a una distancia menor o igual a 20 metros entre sí. En la tabla 6.3 se comprueba esta percepción al ver la distancia que existe entre cada uno de los tuits del usuario. Teniendo calculadas las distancias se puede determinar si el usuario es un *bot* o no. Dicho lo anterior, se puede llegar a la conclusión que el usuario no representa un *bot*.

Para la detección de bots, es preciso señalar el costo computacional que tiene el cálculo de distancia entre tuits de un mismo usuario, sobre todo cuando se cargan miles o millones de datos; eso lo puede convertir en una tarea un tanto prohibitiva en la fase de ETL y tiene que realizarse en un segundo plano y fuera

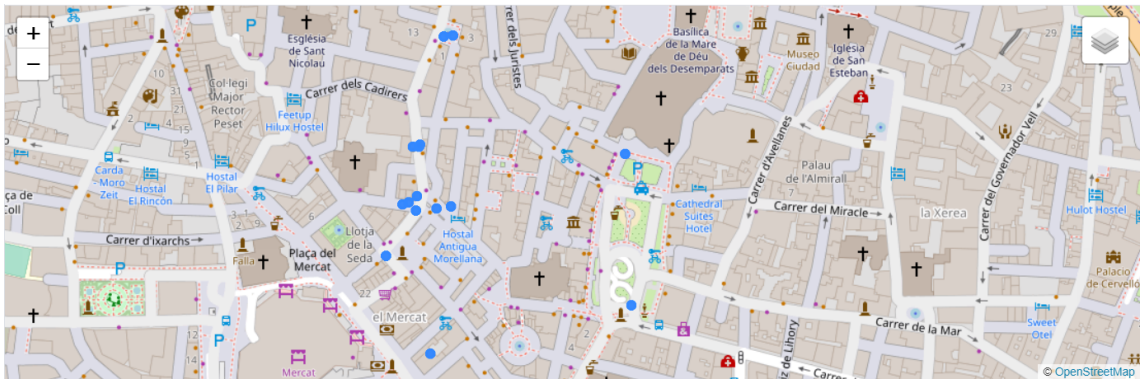


Figura 6.4: Tuits del usuario 1011001099

Tabla 6.3: Ejemplo de distancias entre los tuits del usuario

tuit 1	tuit 2	Distancia
623520953545457664	623520997153636353	50,89 metros
623520953545457664	623520831793139713	100
623520953545457664	623520750385922049	493 metros
623520953545457664	623255340633714688	75,34 metros
...	...	...

de línea. Por otro lado, una limitación que tiene el enfoque seguido es que puede presentarse que un usuario escriba todos sus tuit en la noche al llegar al hotel y sea marcado como un *bot*. No obstante, esta presencia de falsos positivos, con este enfoque se permite excluir un elevado porcentaje de los *bot*.

6.1.3. Procedimiento para la identificación de los turistas

Como ya se ha comentado, todos los análisis en los que se centra BITOUR se relacionan con la actividad de los turistas. Por ello, para el correcto funcionamiento de la plataforma, se debe poder distinguir quiénes son turistas y quiénes no.

Este problema se ha abordado en algunos trabajos, pero hasta la fecha no existe una solución satisfactoria. La fuente de información utilizada para abordar esta tarea es un factor clave en el diseño de un método de identificación de turistas. En nuestro trabajo, la única información disponible son tuits geolocalizados en un destino, es decir, no se usa información disponible en el *timeline*[1] del usuario. En consecuencia, los métodos que se basan en tuits publicados desde diferentes

[1]La línea de tiempo contiene los tuit y retuits realizados por la persona de la cual se está visualizando la información.

países o desde ubicaciones distantes, como [97], no son aplicables en este trabajo. El inconveniente de utilizar las ubicaciones proporcionadas por los usuarios, como en [113], es que esta información no es confiable o incluso es inválida en muchos perfiles de usuario. En general, la mayoría de las soluciones existentes se basan únicamente en la duración del período de publicación del usuario [131, 54], lo cual no es lo suficientemente informativo para discernir entre los residentes y los turistas.

El enfoque utilizado en BITOUR es un poco más sofisticado: está enfocado como la resolución de un problema de reconocimiento de patrones. La idea básica es descubrir patrones en los datos, para que se pueda distinguir entre turistas y residentes. En este caso, los datos no están etiquetados, por lo que es necesario aplicar una técnica de aprendizaje automático no supervisada, en concreto, una técnica de agrupación. Dado un conjunto de características, se obtiene una serie de grupos utilizando el algoritmo *k-means*. Luego, estos grupos se interpretan para decidir cuál de ellos describe a turistas o residentes. A continuación se detalla el proceso.

El primer paso en el problema del reconocimiento de patrones es definir las variables (o características) que se utilizarán en el proceso de agrupamiento. En este caso, inicialmente se seleccionan las siguientes variables que describen el comportamiento de cada usuario. El cálculo del insumo de estas variables se describe en la sección. 4.3.1.4

- Periodo de publicación (`posting_period`): Esta variable también es utilizada en [131, 54]. Se definió un periodo máximo de 30 días.

- Zona horaria del usuario (`time_zone`): esta variable se deriva del idioma del usuario, que determina aproximadamente el país de origen del usuario. Luego, se utiliza la zona horaria de este país, de modo que se pueda calcular una distancia entre tuits con respecto a esta variable.

Variable	Mean	Std	Min	25 %	50 %	75 %	Max
posting_period	9.19	9.35	0.00	1.00	5.00	16.00	30.00
time_zone	1.82	0.79	0.00	1.00	2.00	2.00	9.00
#tweets	13.54	29.19	5.00	6.00	8.00	12.00	796.00
#tagged	4.80	7.26	0.00	1.00	3.00	6.00	98.00
%tagged	44.59	35.22	0.00	9.09	43.47	78.86	100.00
museos	1.23	5.99	0.00	0.00	0.00	0.00	83.33
monumentos	17.33	23.27	0.00	0.00	4.76	28.57	100.00
nocturnos	-	-	-	-	-	-	-
hotel	3.86	12.50	0.00	0.00	0.00	0.00	100.00
gastronomía	5.87	13.22	0.00	0.00	0.00	5.12	100.00
ocio	11.21	18.97	0.00	0.00	0.00	16.66	100.00
transporte	0.68	3.74	0.00	0.00	0.00	0.00	71.42
compras	2.78	8.32	0.00	0.00	0.00	0.00	100.00

Tabla 6.4: Estadísticas del conjunto de datos de Valencia

- Cantidad de tuits (`#tweets`): número total de tuits publicados por este usuario.
- Cantidad de tuits etiquetados (`#tagged`): la cantidad de tuits a los que se les ha asignado un lugar.
- Porcentaje de tuits etiquetados (`%tagged`): relación entre el número de tuits etiquetados y el número de tuits. Esto da una idea de la densidad de tuits publicados desde lugares turísticos por cada usuario.
- Porcentaje de tuits desde lugares clasificados en las categorías de la Tabla 6.1. Este conjunto de variables (`museums`, `monuments`, `night`, `hotel`, `gastronomy`, `leisure`, `transport` and `shopping`) reflejan el tipo de lugares visitados por este usuario.

Una vez seleccionadas las variables, se aplican dos métodos diferentes para determinar el número óptimo de grupos:

- el *método del codo* [84], que consiste en aplicar el agrupamiento de *k-means*, sobre el conjunto de datos para un rango de valores de k, y para cada valor de k calcular la distorsión (suma de errores cuadrados); se traza esta distorsión y el "codo" de este gráfico, donde la distorsión cambia de disminuir

rápidamente a disminuir lentamente, indica el número óptimo de agrupaciones.

- el *método de silueta promedio* [84], donde el concepto de ancho de silueta implica la diferencia entre la estanqueidad dentro del grupo y la separación del resto. Esta puntuación se calcula para cada grupo y el promedio de todos los grupos se calcula para un rango de valores de k. Los valores del ancho de la silueta se encuentran en el rango de —1 a 1, donde un valor cercano a cero significa que la entidad también podría asignarse a otro grupo; un valor cercano a —1 significa que la entidad está mal clasificada y un valor cercano a 1 significa que el conjunto de datos está bien agrupado.

Como ejemplo, se ha aplicado este procedimiento con el conjunto de datos de la ciudad de Valencia. En primer lugar, se calculan las variables indicadas anteriormente. La descripción estadística de estas variables se muestra en la Tabla 6.4. Cuando se aplica el *método del codo* (ver Figura 6.5), se observa que el número óptimo de grupos es 2. Esto se corresponde con el valor más alto de la puntuación de silueta promedio obtenido, 0.44 para dos grupos.

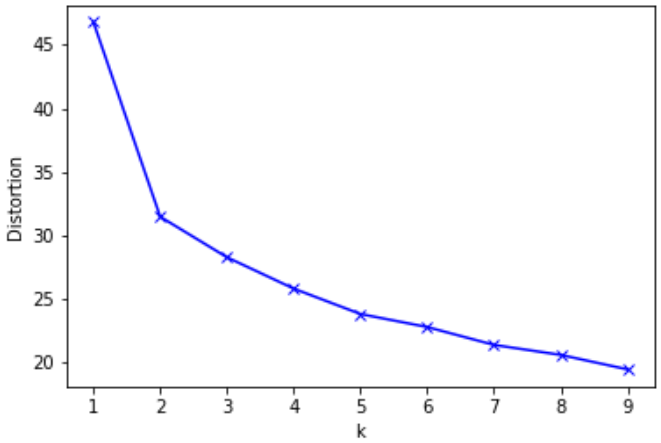


Figura 6.5: Resultado del método del codo para Valencia

La Figura 6.6 muestra un análisis de las variables estudiadas para cada uno de los dos clusters. Se observa que existe una diferencia significativa en el porcentaje de tuits etiquetados por ambos grupos. Además, los usuarios del grupo 0

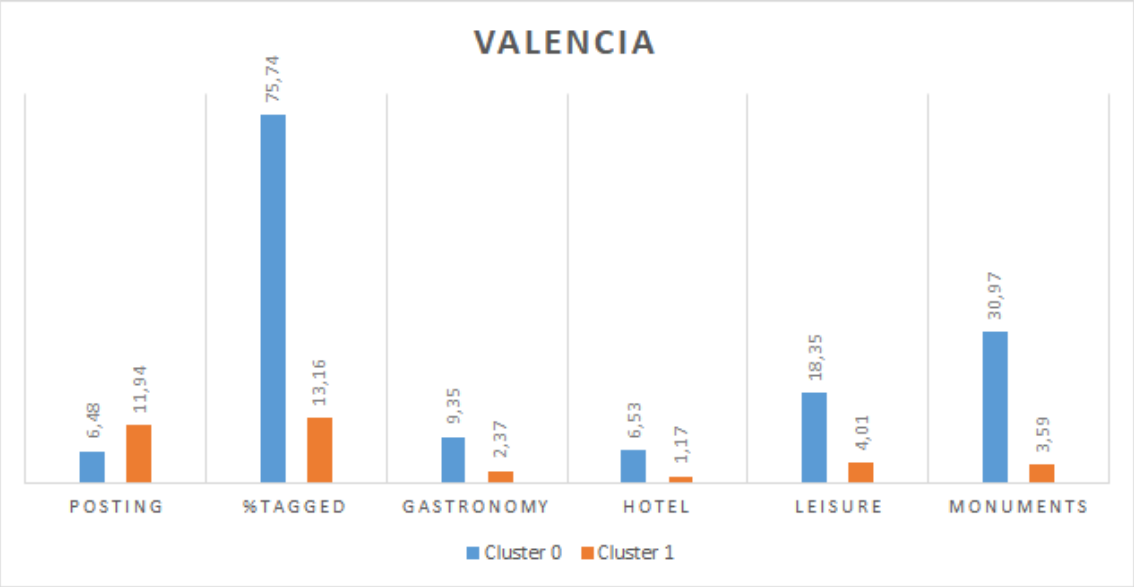


Figura 6.6: Descripción de los grupos para la ciudad de Valencia

presentan un mayor porcentaje de tuits publicados desde cada tipo de ubicación (gastronomía, hotel, ocio y monumentos). En cuanto al período de publicación, el grupo 1 agrupa a los usuarios con un período más largo. Suponemos que los turistas tendrían un período de publicación más corto, dado que permanecen en una ciudad solo temporalmente y publican tuits de lugares turísticos con más frecuencia (y en una mayor densidad) que los residentes. Por estas razones, podemos concluir que el grupo 0 representa a los turistas y el grupo 1 representa a los locales. Esto da como resultado 998 usuarios identificados como turistas en Valencia.

Para validar nuestros resultados, dado que en el proceso de agrupamiento no hemos utilizado ninguna variable relacionada con el idioma del usuario, analizamos el idioma de los usuarios en cada grupo. La figura 6.7 muestra el porcentaje de usuarios en cada grupo (azul para el grupo 0 y naranja para el grupo 1) a los que se les ha asignado cada idioma. En aras de la claridad, solo mostramos los 10 idiomas principales asignados. En la figura se puede observar que la mayoría de los usuarios de habla hispana pertenecen al grupo 1 (locales), mientras que los usuarios asignados a otros idiomas, como el alemán (de), italiano (it), holandés (nl), portugués (pt) o ruso (ru) pertenecen a grupo 0 (turistas).

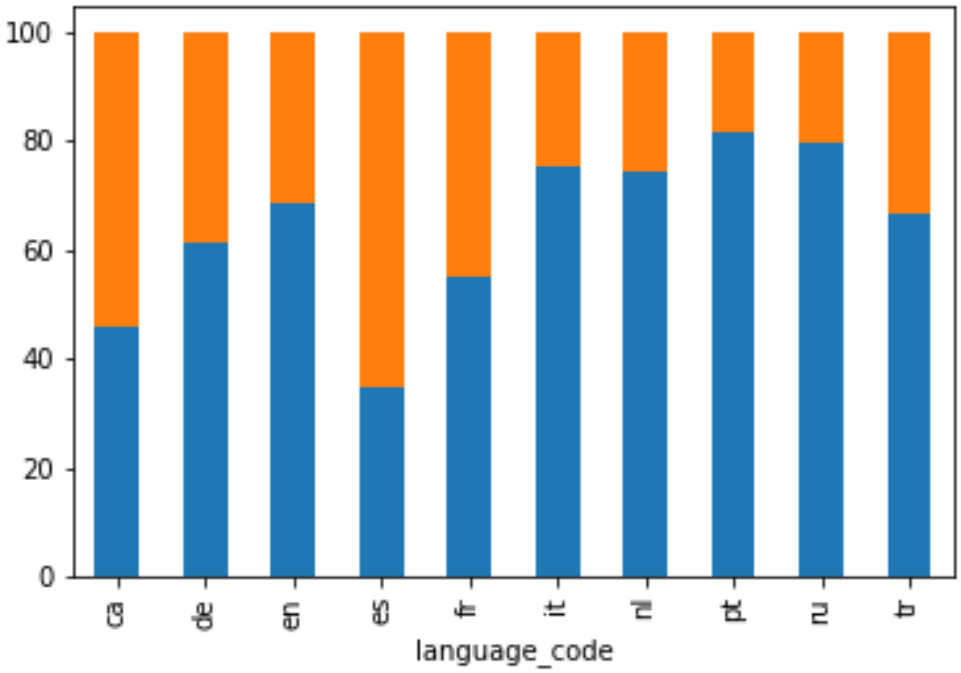


Figura 6.7: Distribución de los 10 idiomas principales en ambos grupos para datos de Valencia

El enfoque seguido para la clasificación de los turistas tiene varias ventajas entre las que se encuentran las siguientes: no depende exclusivamente de la estancia calculada a partir de los tuits (a diferencia de los enfoques existentes); al seguir un enfoque no supervisado posibilita que se vaya afinando a través de la incorporación de nuevas variables; es fácil de entender tanto su funcionamiento como sus resultados. Sin embargo, si se considera que necesita ser validado en más escenarios, y también comparar con usuarios que se sabe son turistas. Otra desventaja es que necesita que el analista determine el significado de cada uno de los grupos (*clusters*) detectados.

6.2 Detalles de la capa de visualización

En esta sección se profundiza con respecto a las formas que ofrece BITOUR para que los usuarios en el rol de analista interactúen con los datos almacenados en la plataforma tras ser procesados, de manera que sirva como soporte para la toma de decisiones en el dominio del turismo. El objetivo es ofrecer los mecanis-

mos necesarios para responder a preguntas tales como: cuáles son las atracciones que mayor afluencia de turistas presentan, qué tipo de atracción (o qué atracción) tiene mayor cantidad comentarios negativos y cuáles lugares prefieren visitar las personas que se hospedan en hoteles.

Esta sección presenta las distintas funcionalidades de BITOUR para el análisis y la visualización de los datos. En primer lugar, se presenta la funcionalidad que ofrece en cuanto a la creación de tablas y gráficos dinámicos en la sección 6.2.1; luego, la visualización de los datos que se puede hacer en un mapa del destino en la sección 6.2.2; y finalmente, en la sección 6.2.3 se aborda la visualización que se puede realizar de la distribución de los tuits alrededor de lugares específicos dentro del destino (por ejemplo, las atracciones más populares).

6.2.1. Tablas y gráficos dinámicos

Esta primera funcionalidad consiste en la posibilidad de crear tablas y gráficos dinámicos a partir de las variables que han sido precargadas. Tanto las tablas como los gráficos son totalmente configurables por el analista mediante las operaciones de arrastrar y soltar variables y los filtros que se pueden aplicar. Tal y como se indica en la figura 6.8, el analista tiene a su disposición una interfaz dividida en cinco secciones:

- **sección 1:** en esta sección se ubican las variables que pueden ser utilizadas para la creación de la tabla dinámica o del gráfico dinámico.

- **sección 2:** en esta sección se selecciona si se desea crear una tabla o un gráfico y, si se desea aplicar alguna función de agregación diferente al conteo, se debe especificar en esta sección, junto a la variable a la cual se le aplica la función.

- **sección 3:** en esta sección se ubican las variables que se desea que formen parte de las columnas para el caso de la tabla mientras que para los gráficos

depende del tipo de gráfico seleccionado, por ejemplo para el diagrama de barra corresponde al eje X.

- **sección 4:** en esta sección se ubican las variables que se desea que formen parte de las filas para el caso de las tablas, mientras que para los gráficos varia en función del tipo de gráfico seleccionado, por ejemplo, para el diagrama de barras corresponde al eje Y.

- **sección 5:** en esta sección se visualiza bien sea una tabla o un gráfico dinámico.

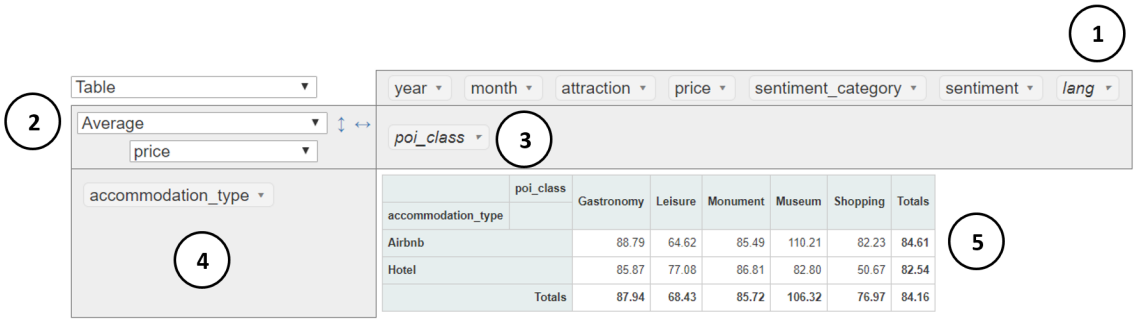


Figura 6.8: Secciones de las tablas o gráficos dinámicos

Teniendo en cuenta la estructura ya descrita, el procedimiento que debe seguir el analista para generar las tablas o gráficos dinámicos es el siguiente:

1. En primer lugar, se debe observar las variables precargadas y que están disponibles para la generación, bien sea de una tabla o de un gráfico dinámico. Por ejemplo, en la figura 6.8 vemos una muestra de las variables que se puede usar:

    - Año (`year`) y Mes (`month`): hace referencia al año y al mes en que los tuits fueron realizados.
    - Tipo de Punto de Interés (`poi_class`): hace referencia a las categorías de las atracciones que fueron asignadas a los tuits.
    - Atracción (`attraction`): representa los nombres de las atracciones del destino y son los lugares a los cuales los tuits son asignados.

- Alojamiento (`accommodation_type`): representa el tipo de alojamiento asignado al usuario que realiza los tuits.
- Costo (`price`): hace referencia al costo de una noche del alojamiento asignado al usuario realizador de los tuits.
- Categoría del sentimiento (`sentiment_category`): representa las categorías asignadas a los tuits basados en su texto, por ejemplo, si el contenido del tuit es de carácter familiar, social, religioso, etc.
- Polaridad del sentimiento (`sentiment`): hace referencia a si el sentimiento que se puede inferir del texto es de carácter positivo o negativo.

2. A continuación, se deben seleccionar las variables para las filas y las columnas de la tabla o ejes del gráfico, por ejemplo, la figura 6.8 muestra que la variable `accommodation_type` es arrastrada para formar parte de las filas de la tabla y la variable `poi_class` de columnas de la tabla.

3. Una vez se tienen las variables seleccionadas, se pueden filtrar los valores de la variable que se desean sean incluidos en la tabla o gráfico. Por ejemplo, de la variable `poi_class` son excluidos los valores transporte (`transport`) y sitios nocturnos (`night`).

4. Finalmente, se debe seleccionar la función de agregación que se utilizará para resumir los datos de la tabla. En la figura 6.8 se aprecia que se seleccionó la función promedio (`average`) y a la variable sobre la que se quiere opere la función de agregación, para el ejemplo, costo (`price`).

A modo de ejemplo, si se desea mostrar la cantidad de tuits realizados en la ciudad de Valencia, España, segregados por el tipo de alojamiento que usaron los turistas (Airbnb u Hoteles) y el tipo del Punto de Interés Turístico (POI) desde el cual se realizó el tuit, se puede tomar la variable `accomodation_type` y arrastrarla a las filas de la tabla; y la variable `POI` se arrastra a las columnas de la tabla. Adicionalmente, si se desea trabajar únicamente con los POI que han sido clasificados como gastronómicos (`gastronomy`), ocio (`leisure`), monumentos

Table | year | month | attraction | sentiment | price | lang | sentiment_category

Count | poi_class

accommodation_type

accommodation_type	poi_class	Gastronomy	Leisure	Monument	Museum	Shopping	Totals
Airbnb		848	476	1,795	202	277	3,598
Hotel		1,622	2,805	3,038	680	342	8,487
Totals		2,470	3,281	4,833	882	619	12,085

Figura 6.9: Configuración de una tabla dinámica

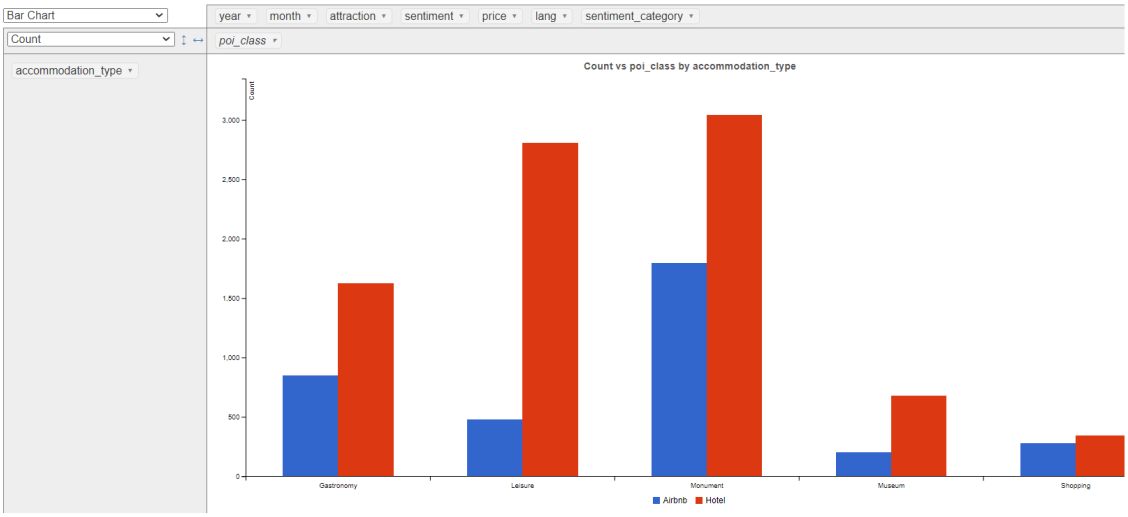


Figura 6.10: Configuración de un gráfico dinámico

(`monuments`), museos (`museums`) y sitios de compras (`shopping`), se puede aplicar un filtro en la variable POI y dejar seleccionados únicamente estos valores. El resultado sería el mostrado en la figura 6.9.

De manera similar se puede operar para crear un gráfico dinámico: se establecen las variables de interés, el tipo de funciones de agregación que se desea utilizar y, en este caso, el tipo de gráfico. En la figura 6.10 se puede ver un gráfico creado a partir de las mismas variables utilizadas para configurar la tabla dinámica. En este caso, se ha seleccionado un gráfico de barras que permite comparar diferentes valores de las variables: los valores de la variable `POI` y los valores de la variable `accomodation_type`. Acá se puede ver cómo los turistas que se hospedan en hoteles (barras rojas) comparados con los que se hospedad en sitios Airbnb (barras azules) realizan mayor cantidad de tuits.

Tanto en el gráfico como en la tabla se aprecia que los turistas que se hospedan en hoteles realizan una mayor cantidad de tuits desde los distintos tipos de POI

siendo los monumentos y los sitios de ocio los tipos de POI los que acaparan la mayor cantidad. Estas diferencias pueden dar un indicio sobre la preferencia de los lugares a visitar de estas dos categorías por los turistas .

Este tipo de análisis de combinación de datos, bien sea mediante tablas o gráficos dinámicos, permite hacer un acercamiento hacia la exploración de las tendencias de los turistas en el destino, por ejemplo, cuáles son las épocas del año preferidas para visitar la ciudad de Valencia.

6.2.2. Filtros y visualización en mapas

Aunque las tablas y gráficos dinámicos constituyen una alternativa valiosa para visualizar los datos, no aprovecha la información espacial de los datos que se están manipulando. Es por esta razón que existe otra alternativa que ofrece BITOUR y es la de visualizar de manera interactiva en un mapa del destino la forma como se concentran los tuits, y filtrar estos por las diferentes variables disponibles. Este mapa, como se aprecia en la figura 6.11 está compuesto por dos partes:

- En la parte superior están dispuestas las variables por las cuales se pueden analizar los tuits que aparecen en el mapa. Es clave aclarar que se pueden seleccionar las variables una independiente de la otra y que además dentro de una variable, se pueden seleccionar aquellos valores que se desean analizar. Estás son las mismas variables descritas en la sección tablas y gráficos dinámicos.

- La parte inferior corresponde al mapa que visualiza los tuits de acuerdo a toda la configuración que se ha realizado previamente en la parte superior con la manipulación de las variables.

Teniendo en cuenta esta estructura, el proceso para generar un mapa con esta alternativa es mucho más sencillo que la anterior.

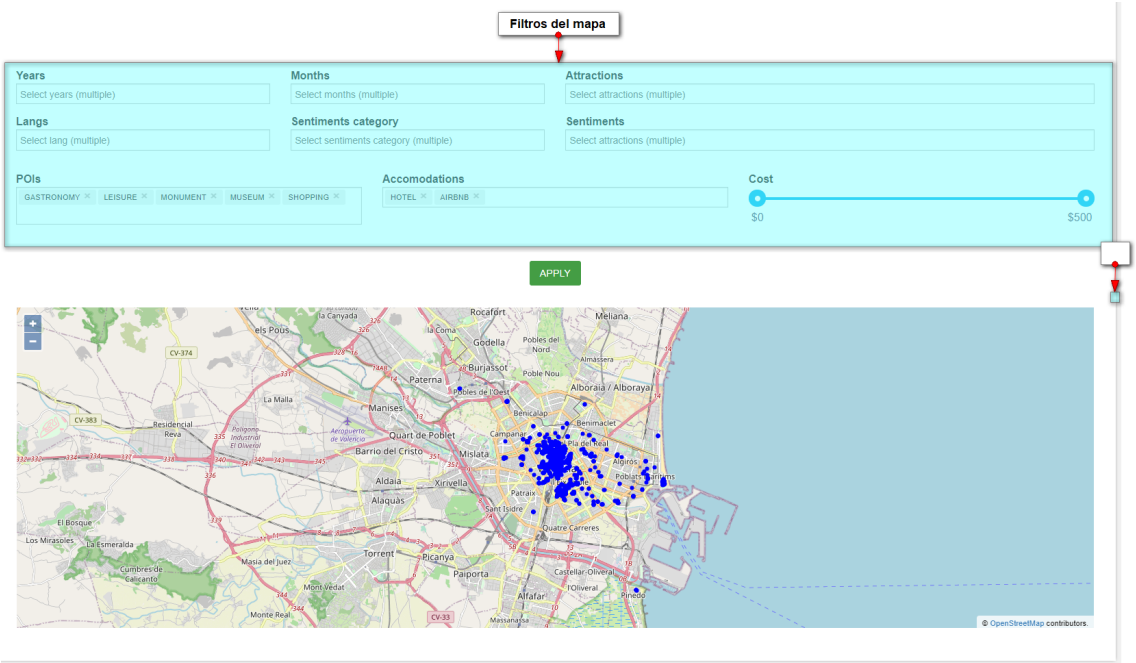


Figura 6.11: Filtros en el mapa

1. El mapa inicia con todos los tuit precargados en el mapa y visualizados como unos puntos de color azul.

2. De todas las variables disponibles en la parte superior se seleccionan los valores de las variables por las cuales se desean filtrar los tuits.

3. Se presiona el botón Aplicar para que en el mapa se reflejen los cambios hechos sobre los valores de las variables.

4. Se puede navegar en el mapa acercándose o alejándose de las zonas visualizadas a discreción.

A modo de ejemplo, si se deseasen seleccionar todos los tuits que fueron realizados en el lenguaje español desde sitios de ocio, se puede utilizar la variable Lenguaje (`Lang`) y seleccionar el valor Español (`spanish`) y en la variable `POIs` y luego seleccionar la variable sitios de ocio (`Leisure`). Esta configuración se puede apreciar en la figura 6.12.

Este tipo de visualización permitiría explorar, por ejemplo, el impacto que tienen los eventos realizados en el destino en cuanto si aumentó el flujo de turistas y la concentración de los mismos. Por ejemplo, en la ciudad de Valencia se puede

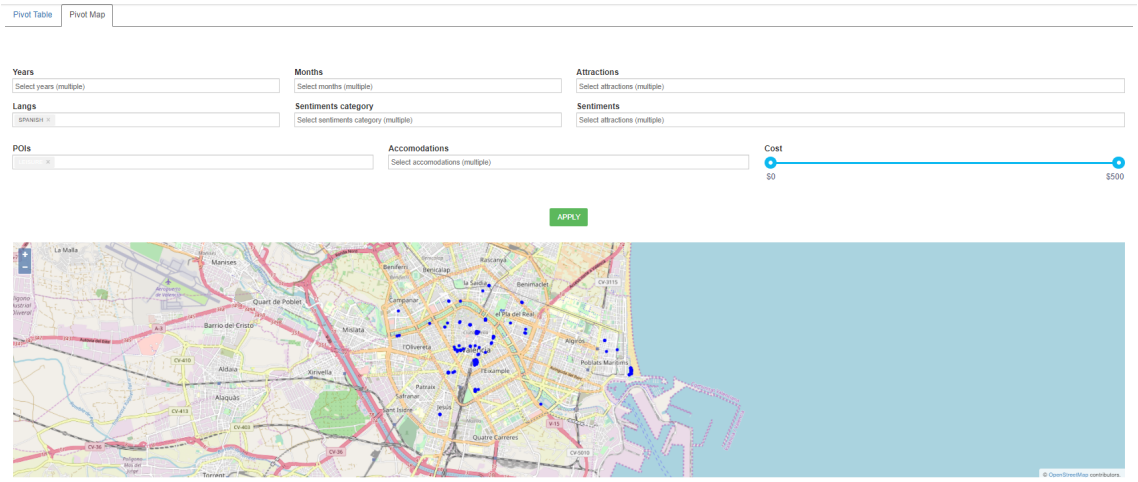


Figura 6.12: Visualización de los tuits en español desde sitios de ocio

analizar durante el periodo de las fallas[2] dónde se concentran principalmente los turistas.

6.2.3. Distribución de tuits alrededor de las atracciones

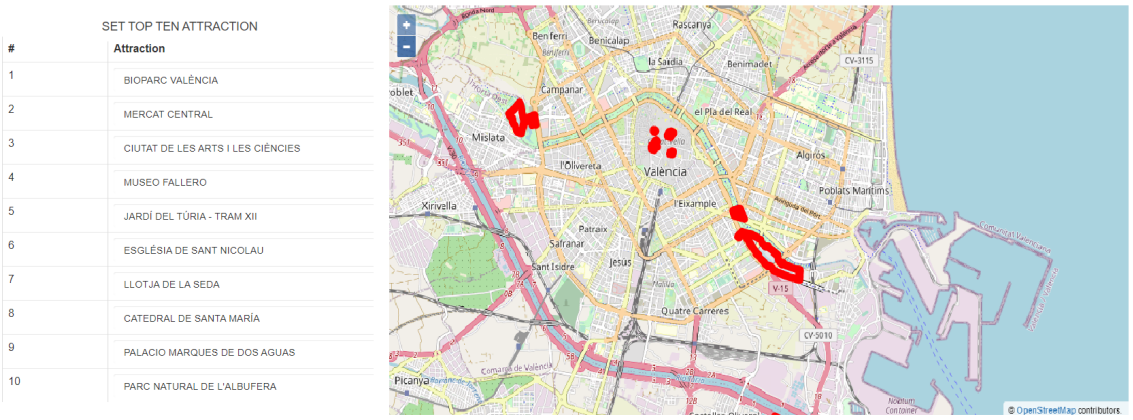
Esta funcionalidad permite visualizar la distribución de los tuits alrededor de lugares de interés para el análisis. La funcionalidad se divide en dos pasos: por un lado, se pueden especificar cuáles son los lugares que se desean analizar y por otro, se puede realizar el análisis de esos lugares definidos.

Para el primer paso, BITOUR proporciona una interfaz como la mostrada en la figura 6.13, en la que el analista puede definir cada uno de los lugares de interés. Esta interfaz se divide de la siguiente forma:

- En la parte izquierda se especifica el nombre de cada una de los lugares a analizar. En la figura se aprecia como se definen diez lugares, entre los que tenemos al Bioparc en el primer lugar.
- En la parte izquierda se visualiza en un mapa todos los lugares que han sido definidos, estos se pintan de color color rojo en el mapa.

Una vez se han definido los lugares, se puede continuar con el segundo paso o de análisis. Para cumplir con este propósito, BITOUR pone a disposición del

[2]Las fallas son una festividad tradicional que tiene lugar en Valencia, España.



SET TOP TEN ATTRACTION

#	Attraction
1	BIOPARC VALÈNCIA
2	MERCAT CENTRAL
3	CIUTAT DE LES ARTS I LES CIÈNCIES
4	MUSEO FALLERO
5	JARDÍ DEL TÚRIA - TRAM XII
6	ESGLÉSIA DE SANT NICOLAU
7	LLOTJA DE LA SEDA
8	CATEDRAL DE SANTA MARÍA
9	PALACIO MARQUES DE DOS AGUAS
10	PARC NATURAL DE L'ALBUFERA

Figura 6.13: Definición de los lugares a analizar

analista una interfaz como la que se aprecia en la figura 6.14, que consta de cuatro zonas de interacción:

- Capas de datos: en esta sección se ubican las capas de datos que es posible agregar o quitar al mapa. Cada capa de datos es un conjunto de tuits clasificados de manera distinta, por ejemplo, se tiene una capa para todos los tuits realizados desde sitios gastronómicos (representados con puntos rosas), una capa para los tuits realizados desde sitios de ocio (representados con puntos azules) y una capa con todos los tuits realizados en el destino (representados con puntos grises).

- Listado de lugares a analizar: en esta sección se ubican todos los lugares que han sido predefinidos como de interés para análisis. Es posible seleccionar cualquiera de las atracciones allí listadas para centrar el análisis en ella. Al seleccionar una de las atracciones el mapa se centra en la ubicación geográfica de la atracción seleccionada y el grupo de estadísticas mostrando cuáles corresponderán a esta atracción.

- Estadísticas básicas del lugar a analizar: en esta sección se muestran estadísticas básicas de los lugares bajo análisis tales como: cantidad de lugares, cantidad de turistas y la cantidad de tuits. Al cargar el mapa la primera vez se mostrarán las estadísticas del destino en su conjunto, pero al seleccionar

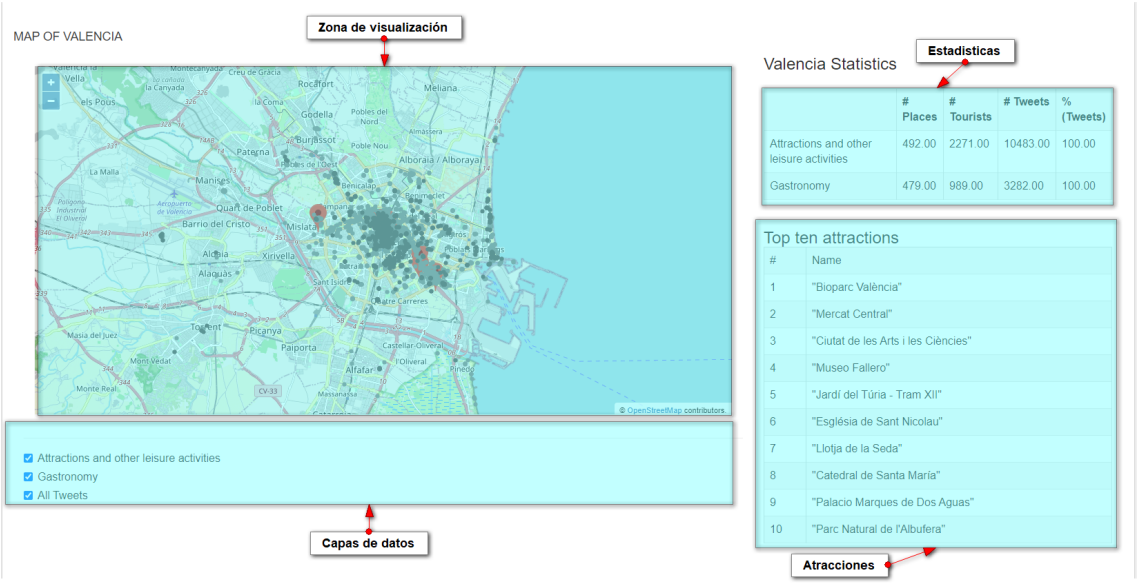


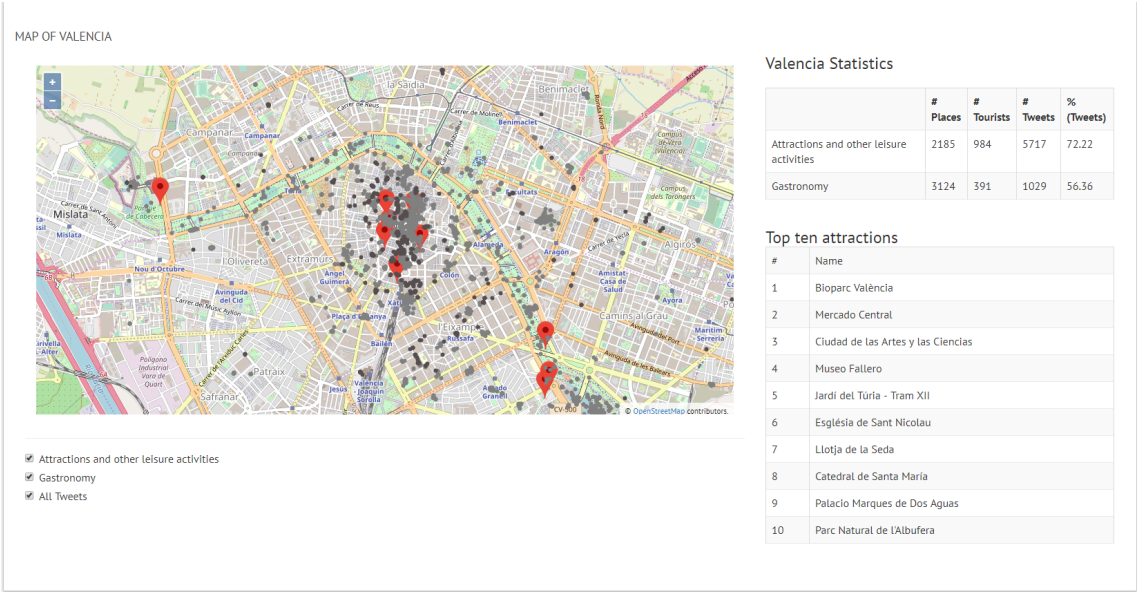
Figura 6.14: Estructura del análisis de distribución de los tuits

cualquiera de las atracciones, las estadísticas corresponderán a la atracción seleccionada.

- Visualización del mapa: en esta sección se combinan todas las opciones anteriores y se despliegan los datos en un mapa del destino bajo análisis. Este mapa permite visualizar los datos para los lugares del listado o para uno en particular, posibilitando además la navegación en el mapa.

A modo de ejemplo, en la figura 6.15 se muestra la distribución de los tuits realizados en la ciudad de Valencia. En esta se muestra que los tuits se encuentran principalmente en los 10 sitios principales de la ciudad de Valencia, mientras otras partes populares de la ciudad, como el área alrededor de la playa, concentran un menor número de tuits y por lo tanto, no se ubica entre los 10 mejores sitios de la ciudad.

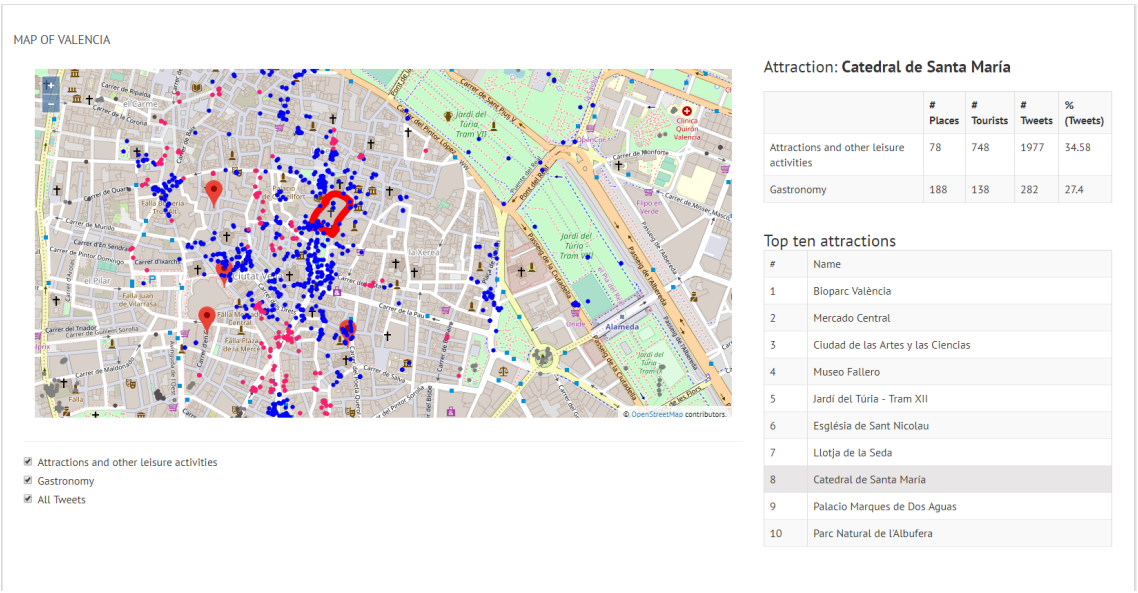
De igual manera, se puede hacer énfasis en una atracción en particular. La figura 6.16 es la vista que muestra la herramienta cuando se selecciona un sitio específico — por ejemplo, la “Catedral de Santa María ” (sitio 8). El mapa destaca el polígono del sitio y los tuits de ambas categorías publicados en un radio de 500 m del sitio (puntos azules para atracciones y puntos rosas para gastronomía).



	# Places	# Tourists	# Tweets	% (Tweets)
Attractions and other leisure activities	2185	984	5717	72.22
Gastronomy	3124	391	1029	56.36

#	Name
1	Bioparc València
2	Mercado Central
3	Ciudad de las Artes y las Ciencias
4	Museo Fallero
5	Jardí del Túria - Tram XII
6	Església de Sant Nicolau
7	Llotja de la Seda
8	Catedral de Santa María
9	Palacio Marques de Dos Aguas
10	Parc Natural de l'Albufera

Figura 6.15: Vista general de destino



	# Places	# Tourists	# Tweets	% (Tweets)
Attractions and other leisure activities	78	748	1977	34.58
Gastronomy	188	138	282	27.4

#	Name
1	Bioparc València
2	Mercado Central
3	Ciudad de las Artes y las Ciencias
4	Museo Fallero
5	Jardí del Túria - Tram XII
6	Església de Sant Nicolau
7	Llotja de la Seda
8	Catedral de Santa María
9	Palacio Marques de Dos Aguas
10	Parc Natural de l'Albufera

Figura 6.16: Catedral de Santa Maria

Este tipo de visualización permite explorar el impacto que tienen unas atracciones en particular, por ejemplo, en atraer a turistas a los sitios aledaños a estas. Por ejemplo, qué tanto jalona a la ciudad de las artes y las ciencias, la visita de los sitios gastronómicos que están a su alrededor.

6.3 Resumen

En este capítulo se presentaron aquellos elementos que se consideran de interés para el lector del documento con respecto a cómo BITOUR ejecuta algunas de sus tareas claves. Es así como se abordó el proceso de asignación de los tuits a los lugares de acuerdo a la distancia que tengan y a la prioridad de la categoría a la que pertenecen los lugares. También se abordó cómo BITOUR a través de técnicas de agrupamiento permite identificar qué usuarios pueden ser considerados como turistas y cuáles no.

Finalmente, se detallaron las alternativas de visualización que proporciona BITOUR para el análisis de los datos. Para esto se inició presentando cómo está conformada la interfaz de cada una de las alternativas visuales: tablas y gráficos dinámicos, filtros y visualización en mapas y distribución de los tuits alrededor de las atracciones. Además, la forma en cómo estas alternativas visuales pueden ser utilizadas para abordar preguntas como: ¿cuáles son las atracciones preferidas por los turistas?, ¿qué percepción tienen del destino?, etc.

CAPÍTULO 7

Conclusiones y trabajo futuros

Como se presentó en la sección 1.2, el objetivo principal de esta tesis de doctorado era desarrollar una solución de Inteligencia de Negocios que incorporara información colaborativa y abierta para soportar la toma de decisiones, tomando como escenario de ejemplificación su aplicación en el sector turístico. Específicamente, la Inteligencia de Negocios se ha utilizado en esta tesis para crear una plataforma denominada BITOUR, que posibilita extraer valor desde los datos disponibles en diferentes fuentes accesibles vía Web. La plataforma recopila de manera automática datos desde diferentes fuentes; los integra en un almacén de datos que los unifica en un formato consistente que permite su posterior procesamiento y la aplicación de técnicas de minería de datos; de manera que luego los datos pueden ser visualizados a través de la web, por la comunidad interesada en analizar los destinos desde una perspectiva turística. A continuación se describen las principales contribuciones de esta tesis:

- El contenido creado de manera colaborativa a través de la Web, como el utilizado en esta tesis, entiéndase aquel donde los usuarios contribuyen en la creación de los datos que una plataforma o sitio Web visualiza, se ha convertido en una fuente de valiosa información que puede ser utilizada en muchos dominios con el propósito de entender el comportamiento y los gustos de las personas (turistas en este caso). En este trabajo se utilizaron en un dominio especifico, el turismo, para analizar la estadía de los turistas en

el destino, los lugares que visitan y la percepción que tienen de éstos. Por lo anterior, una de las principales contribuciones de esta tesis es mostrar cómo las fuentes colaborativas pueden aportar valor al momento de analizar el sector del turismo.

- En el capítulo 2 se construyó el estado del arte con respecto a la Inteligencia de Negocios. Este refleja las diferentes acepciones que se tienen del término, tanto desde una perspectiva administrativa como tecnológica; también, muestra cómo ha evolucionado el término desde estar centrado en el filtrado de información hasta convertirse en un término sombrilla que incorpora una variedad de tecnologías, cada una con un propósito diferente. Igualmente, se identificaron cuáles son las tecnologías que hacen parte del concepto de Inteligencia de Negocios, a saber, los procesos de ETL, el almacén de datos, los cubos OLAP, la minería de datos y las herramientas de visualización; Se expuso que la arquitectura de cuatro capas es la arquitectura típica para construir plataformas de Inteligencia de Negocios y adicionalmente, cómo se articulan cada una de estás tecnologías dentro de la capas de la arquitectura. Finalmente, también se muestra el amplio dominio de aplicaciones que tiene la BI en sectores como el bancario, salud y turismo.

- En la sección 4.2 se describe cómo se diseñó BITOUR, es decir, cuál fue la responsabilidad que se asignó a cada una de las cuatro capas que constituyen la plataforma. Esta arquitectura de cuatro capas permitió articular los componentes que se hacen necesarios para el correcto funcionamiento de la plataforma. También se describe en la sección 4.3 cómo la plataforma se articula alrededor de siete funcionalidades que se encuentran asignadas a los dos roles definidos. Esto es, el rol del administrador, que tiene el propósito de gestionar la creación de los destinos y su configuración, y el rol del analista, pensado para aprovechar todos los datos procesados y los destinos configurados.

- Como se ha descrito en el capítulo 5, se logró cumplir con el objetivo de extraer los datos desde las diferentes fuentes colaborativas que conforman BITOUR: Twitter, OSM, Tripadvisor y Airbnb. El proceso de Extracción, Transformación y Carga para cada una de estas fuentes fue diferente: en el caso de Twitter y OSM, este proceso se apoyó en la API proporcionada por estas plataformas; en el caso de Tripadvisor, se utilizó un enfoque invasivo como es el *webscrapping*; y en el caso de Airbnb, se realizó a través de la lectura de datos estructurados en formato CSV descargados previamente.

- Desde el punto de vista técnico y conceptual, como se describe en las secciones 6.1.1 y 6.1.3, se aportó en introducir un enfoque diferente y con resultados satisfactorios para la asignación de los tuits a los lugares dentro del destino y para la identificación de los turistas de un destino. Los resultados permitieron comprobar que la asignación basada en los criterios de distancia y prioridad y que la identificación de turistas con el algoritmo de agrupamiento ofrecen buenos resultados como en el caso de la ciudad de Valencia. En este caso, el grupo de los usuarios catalogados turistas, contenía características propias de un turista.

- La plataforma creada está diseñada de manera que sea extensible en términos de las fuentes que maneja, es decir, al núcleo base conformado por la información geo-espacial y de opiniones se pueden añadir otras fuentes de un carácter más específico. Esto se evidenció con la incorporación de la información de los hoteles y de los apartamentos turísticos, siendo únicamente necesario que exista una homologación entre los sitios de naturaleza específica y los que aparecen en la cartografía de OSM.

La plataforma BITOUR es especialmente útil para países en vías de desarrollo, como es el caso de Colombia, donde la información disponible para entender a los turistas es escasa. Además de que la inversión en encuestas para recopilar información sobre las actividades que los turistas realizan y de la percepción que

ellos tienen de los destinos es limitada. Esto se debe principalmente a la poca importancia que se le otorga a esta información o, generalmente, por la escasez de recursos financieros de los que disponen estos países para invertir en encuestas u otros estudios de recolección de información. Quedan, de esta manera, en desventaja con respecto a aquellos países que si invierten en este tipo de iniciativas.

7.1 Trabajo futuro

BITOUR se convierte en el primer acercamiento del autor en la utilización de datos colaborativos y de información geográfica que algunos de estos poseen, para entender el comportamiento de los turistas. A partir de esta investigación se desprenden una amplia variedad de temas que merecen ser abordados. Los siguientes párrafos describen algunas líneas de trabajo que se consideran potencialmente interesantes para futuras investigaciones:

- Existen otras fuentes que pueden ser incorporadas para complementar la visión de BITOUR. Algunas de estas fuentes pueden ser *FourSquare* que proporciona detalles sobre los desplazamiento de los usuarios en los destinos e *Instagram* que permite comprender de mejor manera las actividades de ocio que los turistas realizan en el destino.

- No obstante los importantes beneficios que tienen las fuentes de datos colaborativas ya mencionadas y exploradas en la plataforma BITOUR, su utilización también expone nuevos retos debido a que la calidad de estos datos no está garantizada a razón de la libertad que tienen los usuarios de escribir lo que deseen. A causa de esto, y a pesar de que diversos estudios han demostrado que la calidad de estas fuentes se aproximan a la de las fuentes oficiales, sería importante complementar la información colaborativa con datos oficiales y abiertos procedentes de fuentes territoriales, nacionales e internacionales, que ayuden al entendimiento que proporcionan las fuentes colaborativas. Algunas de las fuentes que se deben explorar son los datos

proporcionados por la Organización Mundial del Turismo en su compendio anual sobre estadísticas turísticas de los países y el reporte de la competitividad turística divulgado por el Foro Económico Mundial.

- Otro tipo de fuentes de datos que debe ser explorada para su incorporación dentro de BITOUR son las conocidas como datos enlazados (*Linked Data*), de manera que se pueda navegar de forma automática por la información contenida en la web y así complementar la información que ya se posee, por ejemplo, de las atracciones. Planteado lo anterior, en futuras iteraciones del desarrollo de la plataforma se debe explorar fuentes como DBPedia y LinkedGeoData. La primera debido a que pondría a disposición de BITOUR los datos generados por la enciclopedia web más grande del mundo (Wikipedia); mientras que LinkedGeoData permitirá acceder a la información de OSM de una manera más sencilla y considerando su semántica.

- Desde el punto de vista de la implementación interna de la plataforma hay tres rutinas que pueden ser refinadas:

  - Aunque el algoritmo para la identificación de turistas utilizado en BITOUR va más allá de los métodos tradicionales para la identificación de turistas que se basan exclusivamente en el periodo en el que se realizan los tuits, utilizando para ello técnicas de aprendizaje automático como el agrupamiento y que este algoritmo mostró buenos resultados con las variables utilizadas, se debe explorar la utilización de otras variables relacionadas con el contenido del texto del tuit, como puede ser las referencias que se hacen a las atracciones del destino.

  - El algoritmo empleado para la identificación de los usuarios que son bots es bastante rústico e ingenuo. Este algoritmo está basado únicamente en la proximidad entre todos los tuits que realiza un usuario. Lo anterior, puede desencadenar en la presencia de muchos falsos positivos, por consiguiente, se debe contemplar el enriquecer este algoritmo

empleando más información como la procedente de la línea de tiempo de los usuarios y/o el texto presente en los tuits que él realice.

- El algoritmo de la asignación de los tuits a los lugares puede ser mejorado utilizando otras técnicas de extracción de información desde el texto como lo es la de extracción de entidades, de manera que se pueda conocer a qué entidades (atracciones, museos o monumentos) del destino se hace referencia en el texto de los tuits y utilizar esta información para asignar dichos tuits a los lugares con mayor precisión.

En este sentido el enfoque utilizado tiene otras limitaciones que puede desencadenar en nuevos trabajos, entre las que están:

- Es importante promocionar la herramienta para que sea utilizada por diferentes tipos de usuarios y proporcionar un canal para obtener realimentación de sus comentarios sobre la herramienta, de manera que su opinión sea considerada para la mejora de la herramienta.
- La interfaz de la herramienta actualmente está en el idioma inglés, por lo que es importante incorporar un componente de internacionalización que permita adaptar el contenido diferentes idiomas, inicialmente al español.
- El enfoque actual para la detección del idioma del usuario confía en la detección automática que hace Twitter de cada uno de los tuits realizados en su plataforma. En el futuro se debe probar la utilización de alguna librería especializada para la detección del idioma de cada tuit y comparar estos resultados con los obtenidos actualmente.
- En la herramienta se incluyen únicamente tuits que expresan un sentimiento positivo o negativo (a modo de cantidad), pero sería pertinente explotar el contenido de todos los tuits, incluyendo los de sentimiento neutro, para saber, por ejemplo, qué palabras son las más mencionadas sobre un lugar en particular.

En resumen, a pesar de las limitaciones aquí expuestas, se considera que los objetivos marcados al inicio de la tesis se han cumplido a satisfacción, de manera que se pudo crear un estado del arte alrededor del concepto de Inteligencia de Negocios, se realizó un análisis de fuentes de datos colaborativas que pudiesen ser utilizadas para analizar el sector turismo y se identificaron cuatro fuentes que proporcionan información útil para este análisis; se crearon las rutinas necesarias para la extracción de los datos desde la diferentes fuentes, y se creo una plataforma de BI de cuatro capas que articula todo el proceso: la extracción, la integración, el procesamiento y la visualización de información útil para el análisis.

7.2 Publicaciones

Las dos siguientes secciones listan todas las publicaciones científicas del autor derivadas de este trabajo. La clasificación de los artículos producidos se hace de acuerdo al tipo de publicación: en la primera sección se presentan los artículos que se publicaron o se encuentran en evaluación en revistas del *Science Citation Index* (SCI)[1]; luego, en la segunda sección se listan aquellos artículos publicados en las memorias de conferencias incluidas en el escalafón *Computing Research and Education Association of Australasia* (CORE) [2].

7.2.1. Artículos en revistas del SCI

- Publicado. Alexander Bustamante, Laura Sebastiá, Eva Onaindia. Can Tourist Attractions Boost Other Activities Around? A Data Analysis through Social Networks. En Sensors 2019, 19(11), 2612. `https://doi.org/10.3390/s19112612`.

- Publicado. Alexander Bustamante, Laura Sebastiá, Eva Onaindia. BITOUR: a bussiness intelligence tool for tourism data analysis. En ISPRS Internatio-

[1] `http://ip-science.thomsonreuters.com/cgi-bin/jrnlst/jloptions.cgi?PC=K`
[2] `http://www.core.edu.au/`

nal Journal of Geo-Information 2020, 9(11), 671, https://doi.org/10.3390/ijgi9110671.

- Enviado. Alexander Bustamante, Laura Sebastiá, Eva Onaindia. On the representativeness of OpenStreetMap for tourism analysis. A ISPRS International Journal of Geo-Information. 2020.

7.2.2. Artículos en conferencias CORE

- Publicado. Alexander Bustamante, Laura Sebastiá, Eva Onaindia. Exploratory analysis of representativeness of tourism data in OpenStreetMap. Proceedings of 33 International Business Information Management (33 IBIMA 2019). 10-11 April 2019. Granada, España. pp- 4161-4169.

- Publicado. Jesús Ibañez-Ruiz, Alexander Bustamante, Laura Sebastiá, Eva Onaindia. LinkedDBTour: a Tool to Retrieve Linked Open Data about Tourism Attractions. Proceedings of 31 International Business Information Management (31 IBIMA 2018). 25-26 April 2018. Milan, Italy.

.

Bibliografía

[1] Open linked data and mobile devices as e-tourism tools. a practical approach to collaborative e-learning. *Computers in Human Behavior*, 51:618 – 626, 2015. Computing for Human Learning, Behaviour and Collaboration in the Social and Mobile Networks Era.

[2] Alberto Abelló, José Samos, and Fèlix Saltor. Yam2: a multidimensional conceptual model extending uml. *Information Systems*, 31:541–567, 09 2006.

[3] Bushra Abutahoun, Maisa Alasasfeh, and Salam Fraihat. A framework of business intelligence solution for real estates analysis. pages 1–9, 12 2019.

[4] Muhammad Afzaal and Muhammad Usman. A novel framework for aspect-based opinion classification for tourist places. In *International Conference on Digital Information Management (ICDIM)*, 08 2015.

[5] Alireza Alaei, Susanne Becken, and Bela Stantic. Sentiment analysis in tourism: Capitalizing on big data. *Journal of Travel Research*, 58:004728751774775, 12 2017.

[6] Ahmed Loai Ali and Falko Schmid. Data quality assurance for volunteered geographic information. In Matt Duckham, Edzer Pebesma, Kathleen Stewart, and Andrew U. Frank, editors, *Geographic Information Science*, pages 126–141, Cham, 2014. Springer International Publishing.

[7] Eiman Alothali, Nazar Zaki, Elfadil Mohamed, and Hany Alashwal. Detecting social bots on twitter: A literature review. pages 175–180, 11 2018.

[8] Ricardo J Armas, Desiderio Taño, and Francisco J. García-Rodríguez. Airbnb como nuevo modelo de negocio disruptivo en la empresa turística: un análisis de su potencial competitivo a partir de las opiniones de los usuarios. In *XVIII Congreso AECIT*, 11 2014.

[9] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, page 492–499, USA, 2010. IEEE Computer Society.

[10] Eyad Ayoubi and Shadi Aljawarneh. Challenges and opportunities of adopting business intelligence in smes: collaborative model. *DATA '18: Proceedings of the First International Conference on Data Science, E-learning and Information Systems*, pages 1–5, 10 2018.

[11] Erika Bagambiki. Enterprise data warehouse and business intelligence solution. In *ICEGOV '18: Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance*, pages 665–666, 04 2018.

[12] Rodolfo Baggio and Leonardo Caporarello. Decision support systems in a tourism destination: literature survey and model building. In *Proceedings itAIS-2nd Conference of the Italian chapter of AIS (Association for Information Systems)*, 2005.

[13] Nilanjan Banerjee, Dipanjan Chakraborty, Anupam Joshi, Sumit Mittal, Angshu Rai, and Balaraman Ravindran. Towards analyzing micro-blogs for detection and classification of real-time intentions. In *ICWSM*, 2012.

[14] Sebastian Baumbach, Christoph Rubel, Sheraz Ahmed, and Andreas Dengel. Geospatial customer, competitor and supplier analysis for site selection of supermarkets. In *Proceedings of the 2019 2nd International Conference on Geoinformatics and Data Analysis*, ICGDA 2019, page 110–114, New York, NY, USA, 2019. Association for Computing Machinery.

[15] Yvan Bedard, Sonia Rivest, and Marie-Josée Proulx. Spatial on-line analytical processing (solap): Concepts, architectures and solutions from a geomatics engineering perspective. *Data Warehouses and OLAP: Concepts, Architectures and Solutions*, 01 2006.

[16] Rafael Berlanga, Lisette García-Moya, Victoria Nebot, María José Aramburu, Ismael Sanz, and Dolores María Llidó. Slod-bi: An open data infrastructure for enabling social business intelligence. *Int. J. Data Warehous. Min.*, 11(4):1–28, October 2015.

[17] Donald Berndt, Alan Hevner, and J. Studnicki. Hospital discharge transactions: a data warehouse component. In *the 33rd Annual Hawaii International Conference on System Sciences*, page 10 pp. vol.1, 01 2000.

[18] Eivind Bjørkelund, Thomas H. Burnett, and Kjetil Nørvåg. A study of opinion mining and visualization of hotel reviews. In *Proceedings of the 14th International Conference on Information Integration and Web-Based Applications Services*, IIWAS '12, page 229–238, New York, NY, USA, 2012. Association for Computing Machinery.

[19] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2, 10 2010.

[20] Alicia Hernandez Bonilla, Juana Maria Saucedo Soto, Maria de la Luz Rodriguez Garza, Bernardo Amezcua Nunez, Alicia de la Pena de Leon, and Ailed Samira Bustillos Quezada. Tripadvisor: A Platform That Allows To Explore Experiences And Opinions Of Travelers From The City Of Saltillo, Coahuila, Mexico Tripadvisor Plataforma Que Permite Explorar Experiencias Y Opiniones De. *Revista Internacional Administracion & Finanzas*, 10(2):67–77, 2017.

[21] Maria Borges Tiago, Flávio Tiago, and Francisco Amaral. User-generated content: tourists' profiles on tripadvisor. *International Journal on Strategic Innovative Marketing*, pages 137–147, 01 2014.

[22] Barry Brown. Beyond recommendations: Local review web sites and their impact. *ACM Trans. Comput.-Hum. Interact.*, 19(4), December 2012.

[23] Alexander Bustamante, Sebastia Laura, and Eva Onaindia. On the representativeness of openstreetmap for tourism analysis. Technical report.

[24] Alexander Bustamante, Laura Sebastia, and Eva Onaindia. Can tourist attractions boost other activities around? a data analysis through social networks. *Sensors*, 19(11), 2019.

[25] Sonia Cardona. La inteligencia de negocios y su aplicación en algunas empresas del área metropolitana de medellín, 2005.

[26] Malu Castellanos, Chetan Gupta, Song Wang, Umeshwar Dayal, and Miguel Durazo. A platform for situational awareness in operational bi. *Decis. Support Syst.*, 52(4):869–883, March 2012.

[27] Surajit Chaudhuri, Umeshwar Dayal, and Vivek Narasayya. An overview of business intelligence technology. *Commun. ACM*, 54:88–98, 08 2011.

[28] Hsinchun Chen, Roger H. L. Chiang, and Veda C. Storey. Business intelligence and analytics: From big data to big impact. *MIS Q.*, 36(4):1165–1188, December 2012.

[29] Kuan C. Chen. Decision Support System for Tourism Development: System Dynamics Approach. *Journal of Computer Information systems*, 45(1):104–112, 2004.

[30] Xin Chen, Hoang Vo, and Fusheng Wang. Annotating geographical objects in openstreetmap with geo-tagged social media. In *Proceedings of the 9th ACM SIGSPATIAL Workshop on Location-Based Social Networks*, LBSN16, New York, NY, USA, 2016. Association for Computing Machinery.

[31] Mingming Cheng and Xin Jin. What do airbnb users care about? an analysis of online review comments. *International Journal of Hospitality Management*, 76, 05 2018.

[32] Alton Chua and Snehasish Banerjee. Reliability of reviews on the internet : The case of tripadvisor. In *World Congress on Engineering Computer Science*, 01 2013.

[33] Blazej Ciepluch, Peter Mooney, Ricky Jacob, and Adam C. Winstanley. Using openstreetmap to deliver location-based environmental information in ireland. *SIGSPATIAL Special*, 1(3):17–22, November 2009.

[34] E.F. Codd, S.B. Codd, and C.T. Salley. *Providing OLAP (On-line Analytical Processing) to User-analysts: An IT Mandate*. Codd & Associates, 1993.

[35] L. Cohen. Impacts of business intelligence on population health: a systematic literature review. In *the South African Institute of Computer Scientists and Information Technologists*, pages 1–9, 09 2017.

[36] María del Pilar Salas-Zárate, Estanislao López-López, Rafael Valencia-García, Nathalie Aussenac-Gilles, Ángela Almela, and Giner Alor-Hernández. A study on liwc categories for opinion mining in spanish reviews. *Journal of Information Science*, 40(6):749–760, 2014.

[37] Dursun Delen. *Business Intelligence and Analytics: Systems for Decision Support*. 01 2014.

[38] Nicholas Diakopoulos, Mor Naaman, and Funda Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *VAST 10 - IEEE Conference on Visual Analytics Science and Technology 2010, Proceedings*, pages 115 – 122, 11 2010.

[39] DigitalTOO. Un día de Internet. URL `http://www.digitaltoo.com/`, 2020.

[40] Eckerson. Pervasive business intelligence – techniques and technologies to deploy bi on an enterprise scale, 2009.

[41] Eckerson. Five steps for delivering self-service business intelligence to everyone. techtarget, 2014.

[42] W. W. Eckerson. *Performance dashboards: measuring, monitoring, and managing your business*. John Wiley and Sons, 2 edition, 2010.

[43] Micheline Elias. Enhancing user interaction with business intelligence dashboards. 10 2012.

[44] David W. Embley and Stephen W. Liddle. Big data–conceptual modeling to the rescue. In *Proceedings of the 32nd International Conference on Conceptual Modeling - Volume 8217*, ER 2013, page 1–8, Berlin, Heidelberg, 2013. Springer-Verlag.

[45] Jacinto Estima and Marco Painho. Exploratory analysis of openstreetmap for land use classification. In *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, GEOCROWD '13, page 39–46, New York, NY, USA, 2013. Association for Computing Machinery.

[46] Stephen Few. Information dashboard design : The effective visual communication of data / s. few. 01 2006.

[47] Forrester. Topic overview: Business intelligence, 2008.

[48] World Economic Forum. The travel & tourism competitiveness report. Technical report, World Economic Forum, 2019.

[49] Chiara Francalanci and Ajaz Hussain. Discovering social influencers with network visualization: evidence from the tourism domain. *Information Technology Tourism*, 16, 08 2015.

[50] V. Frias-Martinez, V. Soto, H. Hohwald, and E. Frias-Martinez. Characterizing urban landscapes using geolocated tweets. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 239–248, 2012.

[51] Mohamed Galal, Ghada Hassan, and Mostafa Aref. Developing a personalized multi-dimensional framework using business intelligence techniques

in banking. In *Proceedings of the 10th International Conference on Informatics and Systems*, INFOS '16, page 21–27, New York, NY, USA, 2016. Association for Computing Machinery.

[52] Kevin Gallagher, T. Goles, Stephen Hawk, Judith Simon, Kate Kaiser, C.M. Beath, and Wm Jr. A typology of requisite skills for information technology professionals. pages 1 – 10, 02 2011.

[53] Omar Gambino and Hiram Calvo. A comparison between two spanish sentiment lexicons in the twitter sentiment analysis task. volume 10022, pages 127–138, 11 2016.

[54] Juan Carlos García-Palomares, Javier Gutiérrez, and Carmen Mínguez. Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Applied Geography*, 63:408–417, 2015.

[55] Gartner. Gartner Glossary. URL `https://www.gartner.com/en/`, 2020.

[56] Gartner. Magic quadrant for analytics and business intelligence platforms. Technical report, Gartner, 2020.

[57] Zafar Gilani, Mario Almeida, Reza Farahbakhsh, Liang Wang, and Jon Crowcroft. Stweeler: A framework for twitter bot analysis. 07 2016.

[58] Michael F. Goodchild and Linna Li. Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1:110 – 120, 2012.

[59] Richard Greene. *Business Intelligence and Espionage*. 1966.

[60] Daniel Guttentag. Airbnb: Disruptive innovation and the rise of an informal tourism accommodation sector. *Current Issues in Tourism*, pages 1–26, 12 2013.

[61] Daniel Guttentag. Progress on airbnb: a literature review. *Journal of Hospitality and Tourism Technology*, 10, 06 2019.

[62] Daniel Guttentag, Stephen Smith, Luke Potwarka, and Mark Havitz. Why tourists choose airbnb: A motivation-based segmentation study. *Journal of Travel Research*, 57:004728751769698, 04 2017.

[63] Ido Guy, Avihai Mejer, Alexander Nus, and Fiana Raiber. Extracting and ranking travel tips from user-generated reviews. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 987–996, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[64] Jiawei Han, Micheline Kamber, and Jian Pei. 1 - introduction. In Jiawei Han, Micheline Kamber, and Jian Pei, editors, *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems, pages 1 – 38. Morgan Kaufmann, Boston, third edition edition, 2012.

[65] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41, 11 2013.

[66] Jaime Hernan. La inteligencia de negocios como herramienta para la toma de decisiones estratÉgicas en las empresas. anÁlisis de su aplicabilidad en el contexto corporativo colombiano, 2010.

[67] Wolfram Hopken, Dominic Ernesti, Matthias Fuchs, Kai Kronenberg, and Maria Lexhagen. *Big Data as Input for Predicting Tourist Arrivals*, pages 187–199. 01 2017.

[68] Wolfram Hopken and Matthias Fuchs. Introduction: Special issue on business intelligence and big data in the travel and tourism domain. *Information Technology Tourism*, 16, 03 2016.

[69] Wolfram Hopken, Matthias Fuchs, Gerhard Höll, Dimitri Keil, and Maria Lexhagen. Multi-dimensional data modelling for a tourism destination data warehouse. 04 2013.

[70] Carol Huie. Perceptions of business intelligence professionals about factors related to business intelligence input in decision making. *International Journal of Business Analytics*, 3:1–24, 07 2016.

[71] William Inmon. *Building the Datawarehouse*. John Wiley Sons, Inc., 2002.

[72] Oyku Isik, Mary Jones, and Anna Sidorova. Business intelligence (bi) success and the role of bi capabilities. *Intelligent Systems in Accounting, Finance and Management*, 18:161 – 176, 10 2011.

[73] Mojgan Jadidi, Mir Abolfazl Mostafavi, Yvan Bédard, Bernard Long, and Eve Grenier. Using geospatial business intelligence paradigm to design a multidimensional conceptual model for efficient coastal erosion risk assessment. *Journal of Coastal Conservation*, 17, 09 2013.

[74] Tiaan Jager and Irwin Brown. A descriptive categorized typology of requisite skills for business intelligence professionals. pages 1–10, 09 2016.

[75] Musfira Jilani, Padraig Corcoran, and Michela Bertolotto. Multi-granular street network representation towards quality assessment of openstreetmap data. In *Proceedings of the Sixth ACM SIGSPATIAL International Workshop on Computational Transportation Science*, IWCTS '13, page 19–24, New York, NY, USA, 2013. Association for Computing Machinery.

[76] Jamal Jokar Arsanjani, Peter Mooney, Alexander Zipf, and Marco Helbich. *An introduction to OpenStreetMap in GIScience: Experiences, Research, Applications*. 01 2015.

[77] John Jordan and Clive Ellen. Business need, data and business intelligence. *Journal of Digital Asset Management*, 5, 02 2009.

[78] Nikos Karagiannakis, Giorgos Giannopoulos, Dimitrios Skoutas, and Spiros Athanasiou. Osmrec tool for automatic recommendation of categories on spatial entities in openstreetmap. In *Proceedings of the 9th ACM Conferen-*

ce on Recommender Systems, RecSys '15, page 337–338, New York, NY, USA, 2015. Association for Computing Machinery.

[79] Kijpokin Kasemsap. The fundamentals of business intelligence. *Int. J. Organ. Collect. Intell.*, 6(2):12–25, April 2016.

[80] Mohammed Khan, Muhammad Sohail, Dr. Muhammad Aamir, Bhawani Chowdhry, and Syed Hyder. Web support system for business intelligence in small and medium enterprises. *Wireless Personal Communications*, 76, 03 2014.

[81] Chris Kimble and Giannis Milolidakis. Big data and business intelligence: Debunking the myths. *Global Business and Organizational Excellence*, 35:23–34, 10 2015.

[82] David King. *Business Intelligence: A Managerial Approach*. 01 2008.

[83] Andrew Kipkebut. Structured or unstructured data for deep learning, 02 2020.

[84] Trupti M Kodinariya and Prashant R Makwana. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95, 2013.

[85] Ourania Kounadi. *Assessing the Quality of OpenStreetMap Data*. PhD thesis, 01 2009.

[86] Matthew Krawczyk and Zheng Xiang. Perceptual mapping of hotel brands using online reviews: a text analytics approach. *Information Technology Tourism*, 16, 09 2015.

[87] T. Kuhamanee, N. Talmongkol, K. Chaisuriyakul, W. San-Um, N. Pongpisuttinun, and S. Pongyupinpanich. Sentiment analysis of foreign tourists to bangkok using data mining through online social network. In *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*, pages 1068–1073, 2017.

[88] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? volume 19, 01 2010.

[89] E. Lakomaa and J. Kallberg. Open data as a foundation for innovation: The enabling effect of free public sector information for entrepreneurs. *IEEE Access*, 1:558–563, 2013.

[90] Gregory Lewis and Georgios Zervas. The supply and demand effects of review platforms. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, page 197, New York, NY, USA, 2019. Association for Computing Machinery.

[91] Daming Li, Lianbing Deng, and Zhiming Cai. Statistical analysis of tourist flow in tourist spots based on big data platform and da-hkrvm algorithms. *Personal and Ubiquitous Computing*, 24, 11 2019.

[92] Stephen Litvin and Kaitlyn Dowling. Tripadvisor and hotel consumer brand loyalty. *Current Issues in Tourism*, pages 1–5, 12 2016.

[93] Matthew Love, Charles Boisvert, Elizabeth Uruchrutu, and Ian Ibbotson. Nifty with data: Can a business intelligence analysis sourced from open data form a nifty assignment? pages 344–349, 07 2016.

[94] H. P. Luhn. A business intelligence system. *IBM J. Res. Dev.*, 2(4):314–319, October 1958.

[95] Dennis Luxen and Christian Vetter. Real-time routing with openstreetmap data. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, page 513–516, New York, NY, USA, 2011. Association for Computing Machinery.

[96] Takashi Maeda, Mitsuo Yoshida, Fujio Toriumi, and Hirotada Ohashi. Decision tree analysis of tourists' preferences regarding tourist attractions using geotag data from social media:. pages 61–64, 05 2016.

[97] Takashi Nicholas Maeda, Mitsuo Yoshida, Fujio Toriumi, and Hirotada Ohashi. Extraction of Tourist Destinations and Comparative Analysis of Preferences Between Foreign Tourists and Domestic Tourists on the Basis of Geotagged Social Media Data. *ISPRS International Journal of Geo-Information*, 7(3), 2018.

[98] Elzbieta Malinowski and Esteban Zimanyi. *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications*. 01 2009.

[99] Adam Marcus, Michael Bernstein, Osama Badar, David Karger, Samuel Madden, and Robert Miller. Twitinfo: Aggregating and visualizing microblogs for event exploration. pages 227–236, 05 2011.

[100] Marcello Mariani, R. Baggio, Matthias Fuchs, and Wolfram Höpken. Business intelligence and big data in hospitality and tourism: a systematic literature review. *International Journal of Contemporary Hospitality Management*, 10 2018.

[101] Andrew McAfee and Erik Brynjolfsson. Big data: The management revolution. *Harvard business review*, 90:60–6, 68, 128, 10 2012.

[102] Shah Miah, Huy Vu, John Gammack, and Michael McGrath. A big data analytics method for tourist behaviour analysis. *Information Management*, 54, 12 2016.

[103] Jonathan Milot, Patrick Munroe, Eric Beaudry, Francois Grondin, and Guillaume Bourdeau. Lookupia: An intelligent real estate search engine for finding houses optimally geolocated to reach points of interest. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 651–653, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

[104] Peter Mooney, Padraig Corcoran, and Adam C. Winstanley. Towards quality metrics for openstreetmap. In *Proceedings of the 18th SIGSPATIAL In-*

ternational Conference on Advances in Geographic Information Systems, GIS '10, page 514–517, New York, NY, USA, 2010. Association for Computing Machinery.

[105] Stephen Mooney, Daniel Westreich, and Abdulrahman El-Sayed. Epidemiology in the era of big data. *Epidemiology (Cambridge, Mass.)*, 26, 03 2015.

[106] George Musa, Po-Huang Chiang, Tyler Sylk, Rachel Bavley, William Keating, Bereketab Lakew, Hui-Chen Tsou, and Christina Hoven. Use of gis mapping as a public health tool—from cholera to cancer. *Health services insights*, 6:111–6, 11 2013.

[107] United Nations. World economic situation and prospect, 2019.

[108] Solomon Negash and Paul Gray. Business intelligence. volume 13, page 423, 01 2003.

[109] Nicholas Nicoli and Evgenia Papadopoulou. Tripadvisor and reputation: a case study of the hotel industry in cyprus. *EuroMed Journal of Business*, 12:00–00, 07 2017.

[110] Onook Oh, Manish Agrawal, and H. Raghav Rao. Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Q.*, 37(2):407–426, June 2013.

[111] Babajide Osatuyi. Information sharing on social media sites. *Comput. Hum. Behav.*, 29(6):2622–2631, November 2013.

[112] Mohamed Oubezza and Jamal Elkafi. An approach for the implementation of semantic big data analytics in the social business intelligence process on distributed environments (cloud computing). pages 1–6, 10 2019.

[113] Jose J Padilla, Hamdi Kavak, Christopher J Lynch, Ross J Gore, and Saikou Y Diallo. Temporal and spatiotemporal investigation of tourist attraction visit sentiment on Twitter. *PloS one*, 13(6):e0198857, 2018.

[114] Kostas Pantazos, Mohammad Amin Kuhail, Soren Lauesen, and Shangjin Xu. uvis studio: An integrated development environment for visualization. volume 8654, 02 2013.

[115] Maria C. Papoutsoglou, Nikolaos Vesyropoulos, and Christos K. Georgiadis. Utilizing business intelligence and social media streams for optimized web service compositions. In *Proceedings of the 7th Balkan Conference on Informatics Conference*, BCI '15, New York, NY, USA, 2015. Association for Computing Machinery.

[116] Martijn Paulussen. *Smart Self-Service Business Intelligence Framework*. PhD thesis, 11 2018.

[117] T. B. Pedersen and C. S. Jensen. Multidimensional database technology. *Computer*, 34(12):40–46, 2001.

[118] Michael Peng, Sheng-Hwa Tuan, and Feng-Chi Liu. Establishment of business intelligence and big data analysis for higher education. pages 121–125, 07 2017.

[119] Daniel Power. A brief history of decision support systems. *COM, World Wide Web, http://DSSResources. COM/history/dsshistory. html, version*, 2, 01 2007.

[120] Davide Proserpio and Georgios Zervas. Online reputation management: Estimating the impact of management responses on consumer reviews. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, EC '15, page 79, New York, NY, USA, 2015. Association for Computing Machinery.

[121] D. Punjani, K. Singh, A. Both, M. Koubarakis, I. Angelidis, K. Bereta, T. Beris, D. Bilidas, T. Ioannidis, N. Karalis, C. Lange, D. Pantazi, C. Papaloukas, and G. Stamoulis. Template-based question answering over linked geospatial data. In *Proceedings of the 12th Workshop on Geographic Information*

Retrieval, GIR'18, New York, NY, USA, 2018. Association for Computing Machinery.

[122] Raconteur. A Day in Data. URL `https://www.raconteur.net/infographics/a-day-in-data`, 2019.

[123] Gunupudi Rajesh Kumar, Vangipuram Radhakrishna, and Shadi Aljawarneh. Strategic application of software process model to optimize business intelligence results. 09 2015.

[124] Frederik Ramm, Jochen Topf, and Steve Chilton. *OpenStreetMap:Using and Enhancing the Free Map of the World*. UIT Cambridge, Germany, 2010.

[125] Gregory Richards, G. Yeoh, Alain Chong, and Aleš Popovič. Business intelligence effectiveness and corporate performance management: An empirical analysis. *Journal of Computer Information Systems*, 59:188–196, 01 2019.

[126] Seyed Rizi and Abdul Roudsari. Development of a public health reporting data warehouse: Lessons learned. *Studies in health technology and informatics*, 192:861–5, 08 2013.

[127] Stefano Rizzi. *Collaborative Business Intelligence*, pages 186–205. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[128] M. Sabou, Adrian Brasoveanu, and Irem Onder. Linked data for cross-domain decision-making in tourism. *Journal of Internet Services and Applications*, 6(1):1–13, 2015.

[129] Marta Sabou, Irem Onder, Adrian Brasoveanu, and Arno Scharl. Towards cross-domain decision making in tourism: A linked data based approach. *SSRN Electronic Journal*, 01 2015.

[130] Krit Salah-ddine, Hicham el Bousty, elasikri, M. Kabrane, Kaoutar Bendaoud, Khaoula Karimi, and Hassan Oudani. Investigating business intelligence in the era of big data: Concepts, benefits and challenges. *International Journal of Engineering Science*, 03 2018.

[131] Maria Henar Salas-Olmedo, Borja Moya-Gomez, Juan Carlos Garcia-Palomares, and Javier Gutierrez. Tourists' digital footprint in cities: Comparing Big Data sources. *Tourism Management*, 66:13 – 25, 2018.

[132] Arvind Satyanarayan and Jeffrey Heer. Lyra: An interactive visualization design environment. *Computer Graphics Forum*, 33, 06 2014.

[133] MohammadBagher Sayedi, Peyman Ghafari, and Elham Hojati. Analysis of the effects and factors of implementing cloud business intelligence in banking systems. pages 197–198, 12 2017.

[134] Brenda Scholtz, Martin Smuts, and André Calitz. Design guidelines for business intelligence tools for novice users. 09 2015.

[135] Sukhjit Sehra, Jaiteg Singh, and Hardeep Rai. A systematic study of openstreetmap data quality assessment. pages 377–381, 04 2014.

[136] Parvaneh Shayegh and Negin Daneshpour. Using a data warehouse to improve analyzing tourism data. 04 2015.

[137] Doris Silbernagl, Nikolaus Krismer, and Günther Specht. Comparing osm area-boundary data to dbpedia. In *Proceedings of the 12th International Symposium on Open Collaboration*, OpenSym '16, New York, NY, USA, 2016. Association for Computing Machinery.

[138] Martin Smuts, Brenda Scholtz, and Andre Calitz. Design guidelines for business intelligence tools for novice users. In *Proceedings of the 2015 Annual Research Conference on South African Institute of Computer Scientists and Information Technologists*, SAICSIT '15, New York, NY, USA, 2015. Association for Computing Machinery.

[139] Dario Stojanovski, Ivica Dimitrovski, and Gjorgji Madjarov. Tweetviz: Twitter data visualization. In *Conference on Data Mining and Data Warehouses (SiKDD 2014) - Information Society*, 10 2014.

[140] Tourism Studies, Pablo Limberger, Francisco Anjos, Jéssica Meira, and Sara Anjos. Satisfaction in hospitality on tripadvisor.com: An analysis of the correlation between evaluation criteria and overall satisfaction. *Tourism Management Studies*, 10:59–65, 01 2014.

[141] Arjun Talwar, Radu Jurca, and Boi Faltings. Understanding user behavior in online feedback reporting. In *Proceedings of the 8th ACM Conference on Electronic Commerce*, EC '07, page 134–142, New York, NY, USA, 2007. Association for Computing Machinery.

[142] Mike Thelwall. *Sentiment Analysis for Tourism*, pages 87–104. Springer Singapore, Singapore, 2019.

[143] Tripadvisor. Tripadvisor. URL `https://tripadvisor.mediaroom.com/`, 2020.

[144] Juan Trujillo, Manuel Sanz, Jaime Gomez, and Il-Yeol Song. Designing data warehouses with oo conceptual models. *Computer*, 34:66–75, 01 2002.

[145] Twitter. Q3 2019 letter to shareholders, 2019.

[146] Thanathorn Vajirakachorn and Jongsawas Chongwatpol. Application of business intelligence in the tourism industry: A case study of a local food festival in thailand. *Tourism Management Perspectives*, 23:75 – 86, 2017.

[147] Thanathorn Vajirakachorn and Jongsawas Chongwatpol. Application of business intelligence in the tourism industry: A case study of a local food festival in thailand. *Tourism Management Perspectives*, 23:75 – 86, 2017.

[148] Panos Vassiliadis, Alkis Simitsis, and Spiros Skiadopoulos. Conceptual modeling for etl processes. 12 2002.

[149] Quoc Duy Vo, Jaya Thomas, Shinyoung Cho, Pradipta De, and Bong Jun Choi. Next generation business intelligence and analytics. In *Proceedings of the 2nd International Conference on Business and Information Management*,

ICBIM '18, page 163–168, New York, NY, USA, 2018. Association for Computing Machinery.

[150] Gottfried Vossen and Stephan Hagemann. *Unleashing Web 2.0: From Concepts to Creativity*, volume 2007. 12 2007.

[151] David Wachsmuth and Alexander Weisler. Airbnb and the rent gap: Gentrification through the sharing economy. *Environment and Planning A: Economy and Space*, 02 2018.

[152] Parth Wazurkar, Robin Bhadoria, and Dhananjai Bajpai. Predictive analytics in data science for business intelligence solutions. pages 367–370, 11 2017.

[153] Mary Wisniewski, Piotr Kieszkowski, Brandon Zagorski, William Trick, Michael Sommers, and Robert Weinstein. Development of a clinical data warehouse for hospital infection control. *Journal of the American Medical Informatics Association : JAMIA*, 10:454–62, 09 2003.

[154] Karl W Wöber. Information supply in tourism management by marketing decision support systems. *Tourism Management*, 24(3):241 – 255, 2003.

[155] World Economic Forum. Travel and Tourism Competitiveness Report 2017. URL `http://reports.weforum.org/travel-and-tourism-competitiveness-report-2017/`, 2017.

[156] World Tourism Organization. *Compendium of Tourism Statistics*, volume 2016 Edition. UNWTO, Madrid, Spain, 2016.

[157] Junjie Wu, Haoyan Sun, and Yong Tan. Social media research: A review. *Journal of Systems Science and Systems Engineering*, 22:257–282, 09 2013.

[158] Karen Xie, Chih-Chien Chen, and Shinyi Wu. Online consumer review factors affecting offline hotel popularity: Evidence from tripadvisor. *Journal of Travel Tourism Marketing*, 33:1–13, 07 2015.

[159] Taha Yasseri, Giovanni Quattrone, and Afra Mashhadi. Temporal analysis of activity patterns of editors in collaborative mapping project of openstreetmap. In *Proceedings of the 9th International Symposium on Open Collaboration*, WikiSym '13, New York, NY, USA, 2013. Association for Computing Machinery.

[160] Kyung-Hyan Yoo, Marianna Sigala, and Ulrike Gretzel. *Exploring TripAdvisor*, pages 239–255. 01 2016.

[161] Benxiang Zeng. Social media in tourism. *Journal of Tourism Hospitality*, 2:1–2, 01 2013.

[162] Leishi Zhang, Andreas Stoffel, Michael Behrisch, Sebastian Mittelstädt, Tobias Schreck, Rene Pompl, Stefan Weber, Holger Last, and Daniel Keim. Visual analytics for the big data era — a comparative review of state-of-the-art commercial systems. pages 173–182, 10 2012.

[163] Paul Zikopoulos, Chris Eaton, and IBM. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Osborne Media, 1st edition, 2011.

[164] Esteban Zimányi, editor. *Business Intelligence and Big Data - 7th European Summer School, eBISS 2017, Bruxelles, Belgium, July 2-7, 2017, Tutorial Lectures*, volume 324 of *Lecture Notes in Business Information Processing*. Springer, 2018.

[165] Matthew Zook and Jessica Breen. *Volunteered Geographic Information*, pages 2434–2438. Springer International Publishing, Cham, 2017.