

**UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
FACULTAD DE INGENIERÍA
MAESTRÍA Y DOCTORADO EN CIENCIAS E INGENIERÍA**



**"LOCALIZACIÓN EN INTERIORES VÍA COLABORACIÓN ABIERTA
DISTRIBUIDA Y CAMPO MAGNÉTICO"**

**TESIS PARA OBTENER EL GRADO DE MAESTRO EN CIENCIAS QUE
PRESENTA**

ISABEL LEBASI AMBRIZ SILVA

DIRECTOR

DR. JUAN PABLO GARCÍA VÁZQUEZ

CO-DIRECTOR

DR. ÁNGEL GABRIEL ANDRADE REÁTIGA

Mexicali, Baja California

06 de Marzo de 2018

TESIS DEFENDIDA POR

Isabel Lebasi Ambriz Silva

Y APROBADA POR EL SIGUIENTE COMITÉ

Dr. Juan Pablo García Vázquez

Dr. Ángel Gabriel Andrade Reátiga

Dra. Margarita Gil Samaniego Ramos

Dr. Arnoldo Díaz Ramírez

Dr. Guillermo Galaviz Yañez
Coordinador de Posgrado e Investigación
Facultad de Ingeniería

Mexicali, Baja California,

06 de Marzo de 2018

RESUMEN de la Tesis de Isabel Lebasi Ambriz Silva, presentada como requisito parcial para la obtención del grado de MAESTRO EN CIENCIAS en el campo del conocimiento de COMPUTACIÓN. Mexicali, Baja California, México. 06 de Marzo de 2018.

LOCALIZACIÓN EN INTERIORES VÍA COLABORACIÓN ABIERTA DISTRIBUIDA Y CAMPO MAGNÉTICO

Dr. Juan Pablo García Vázquez
Director de Tesis

Dr. Ángel Gabriel Andrade Reátiga
Co-Director de Tesis

RESUMEN

En esta tesis se describe el experimento realizado para determinar la viabilidad de utilizar la colaboración abierta distribuida como etapa de recolección de datos en la técnica de huella digital de señal. El experimento consistió en la creación de mapa radio con mediciones de intensidad de campo magnético recolectadas en un entorno de interiores compuesto por varias habitaciones. Las mediciones de intensidad de campo magnético fueron utilizadas para generar modelos predictivos para estimar la localización de un individuo en el entorno de interiores. Para la creación de los modelos se utilizaron tres diferentes algoritmos de aprendizaje automático: k vecinos más cercanos (k-NN del inglés k Nearest Neighbors), árboles de decisión (J48) y bayesiano ingenuo (del inglés Naïve Bayes). El rendimiento de los algoritmos se midió a través se consideraron diversas métricas que son obtenidas a partir de la matriz de confusión, estos fueron la exactitud y coeficiente de concordancia (*kappa*). El modelo predictivo con mayor desempeño fue el obtenido con el algoritmo k-NN con un valor de $k=3$, ya que este ofrece una exactitud de 97.12 % para estimar la localización del individuo. Mientras que el modelo con más bajo rendimiento fue el bayesiano ingenuo quien ofrece una exactitud de 50.79 % con un nivel de concordancia (*kappa*) de 0.3834. Los resultados anteriores indican que los datos de campo magnético recolectados mediante la colaboración abierta distribuida permiten

generar un modelo predictivo para estimar la localización de un individuo en un entorno de interiores. Por lo que es viable utilizar la colaboración abierta distribuida para la construcción de sistemas de localización para interiores.

Palabras clave: Aprendizaje Automático, Localización en Interiores, Localización Simbólica, Mapa de Radio, Colaboración Abierta Distribuida y Huella digital de Señal.

ABSTRACT of the thesis presented by Isabel Lebasi Ambriz Silva, in order to obtain the MASTER degree in COMPUTER SCIENCE. Mexicali, Baja California, México. March 6, 2018.

INDOOR LOCALIZATION CONSIDERING CROWDSOURCING AND MAGNETIC FIELD

Dr. Juan Pablo García Vázquez
Thesis Advisor

Dr. Ángel Gabriel Andrade Reátiga
Thesis Co-Advisor

ABSTRACT

This thesis describes the experiment carried out to determine the viability of using distributed open collaboration as a data collection stage in the fingerprint technique. The experiment consisted in the creation of a radio map with measurements of magnetic field strength collected in an indoor environment composed of several rooms. Magnetic field strength measurements were used to generate predictive models to estimate the location of an individual in the indoor environment. For the creation of the models, three different machine learning algorithms were used: k nearest neighbors (k-NN), decision trees (J48) and naïve Bayesian (Naïve Bayes). The performance of the algorithms was measured through the accuracy and the Kappa coefficient. The predictive model with the highest performance was obtained with the k-NN algorithm with a value of $k = 3$, since it offers an accuracy of 97.12% to estimate the location of the individual. The model with the lowest performance was the naïve Bayesian who offers an accuracy of 50.79% with a level of agreement (kappa) of 0.3834. The above results indicate that the magnetic field data collected through distributed open collaboration allow generating a predictive model to estimate the location of an individual in an indoor environment. So it is feasible to use distributed open collaboration for the construction of indoor location systems.

Keywords: Machine Learning, Indoor Localization, Symbolic Localization, Radio Map, Fingerprint, Crowdsourcing

Dedicatoria

A mis padres por el apoyo y motivación a lo largo de mi vida para realizarme profesional y exitosamente.

A mis hermanos Pryscila, Laura y Chillon's por inspirarme a seguir mis metas en la vida.

A mi sobrina Camila por la motivación que me ha brindado durante el logro de mis estudios.

A mi novio Edmundo por el apoyo incondicional en esos días difíciles y de estrés.

A mis amigos César y Margarita por las palabras de apoyo y motivación a lo largo del camino.

Agradecimientos

Quisiera agradecer a:

A mi familia por creer en mí y motivarme a ser una persona exitosa.

A mi director el Dr. Juan Pablo García Vázquez por la orientación, aprendizaje y paciencia brindada a lo largo de mi estudio de maestría.

A mi co-Director el Dr. Ángel Andrade Reátiga por su disponibilidad, apoyo y orientación.

A mi comité de tesis, la Dra. Margarita Gil Samaniego por su disponibilidad y asesoría; y al Dr. Arnoldo Díaz por sus consejos, disponibilidad, asesoría y por apoyarnos con equipos para la experimentación.

Al M.C. Luis Vizcarra y a la M.C. Mónica Lam por la beca y apoyo de transporte para ir a CICOMP 2016.

A los estudiantes del programa educativo de LSC que participaron en la recolección de datos.

Al consejo Nacional de Ciencia y Tecnología (CONACyT) por la beca otorgada para la realización de mis estudios de maestría.

A la Universidad Autónoma de Baja California y a mis compañeros Ricardo y Alexander que me brindaron su apoyo incondicional.

Gracias a todas las personas que ayudaron directa e indirectamente en la realización de mis estudios de maestría.

Isabel Lebasi (firma)

Tabla de Contenido

1

1. Introducción	1
1.1 Motivación y contexto del problema	1
1.2 Colaboración Abierta Distribuida	3
1.3 Objetivos de la Tesis	5
1.3.1 Objetivo General	5
1.3.2 Objetivos Específicos	5
1.4 Metodología	5
1.4.1 Recolección de datos	6
1.4.2 Procesamiento de los datos	6
1.4.3 Generación del modelo de localización	7
1.4.4 Validación del modelo de localización	7
1.5 Organización de la Tesis	8

2

2. Trabajo Relacionado	9
2.1 Introducción	9
2.2 Sistemas para la construcción automática de planos	9
2.2.1 CrowdInside	10
2.2.2 Piloc	12
2.2.3 CIMLoc	13
2.2.4 Groping	14
2.3 Eliminación de la etapa de calibración	15
2.3.1 Freeloc	15
2.3.2 Zee	16
2.3.3 MaWi	17
2.3.4 MaLoc	17
2.4 Conclusiones	18

3

3. Recolección y procesamiento de datos	19
3.1 Recolección de datos	19
3.2 Descripción del entorno de experimentación	20
3.3 Técnica para la recolección de datos	21
3.4 Preprocesamiento de la información	23
3.5 Resultados	24

4

4. Generación del Modelo de Localización	27
4.1. Introducción	27
4.1.1 K vecinos más cercanos	27
4.1.2 Árboles de decisión	29
4.1.3. Naïve Bayes	32
4.2 Implementación de los Algoritmos de Aprendizaje Automático	33
4.3 Descripción del entorno de experimentación	21
4.3 Validación del mapa de radio	33
4.4 Resultados	34

5. Conclusiones y Trabajo Futuro	40
5.1. Conclusiones	40
5.2. Contribuciones	41
5.2.1 Publicaciones de la Tesis	41
5.3 Trabajo Futuro	43
5.3.1 Modelar la señal de campo magnético a través de características temporales, de frecuencia y estadísticas.....	43
5.3.2 Generar un modelo predictivo para entornos de interiores multi-nivel	43
5.3.3 Tamaño de los salones.....	43
5.3.4 Identificar las variaciones del campo magnético en el entorno de interiores por objetos presentes en el entorno	44
A	
ABSTRACT	v
Agradecimientos.....	viii
D	
Dedicatoria	vii
I	
Índice de Figuras.....	xi
Índice de Tablas.....	xii
Introducción	1
R	
Referencias Bibliográficas	45
RESUMEN	iii

Índice de Figuras

Figura 1. Entorno de interiores subdivido en regiones, en las cuales se mide y registra una señal considerando la distancia y orientación.....	2
Figura 2. Nota de usuario que informa que el establecimiento comercial marcado en el mapa ya no existe	4
Figura 3. Metodología de Investigación	6
Figura 4. Rutas generadas con el sistema CrowdInside a partir del desplazamiento de un individuos dentro de un edificio (Fuente: Alzantot et al. 2012)	12
Figura 5. Rutas de un individuo asociadas a un punto de acceso (Fuente: Luo et al., 2014)	13
Figura 6. Mapa digital de un entorno de interiores construido con el sistema CIMloc (Fuente: Zhang et al. 2013)	14
Figura 7. Construcción del plano de un edificio a partir de las trayectorias realizadas por de los individuos en el Proyecto GROPING (Fuente: Zhang et al. 2015).....	15
Figura 8. Funcionamiento del Sistema Freeloc (Fuente: Yang et al., 2013)	16
Figura 9. Ejemplo del Sistema Zee (Fuente: Rai et al., 2012)	17
Figura 10. Interfaz del Sistema MaLoc	18
Figura 11. Interfaz de usuario de la Aplicación	19
Figura 12. Estructura del documento <i>JSON</i> que almacena las mediciones	20
Figura 13. Plano del Laboratorio de Sistemas Computacionales.....	21
Figura 14. Invitación realizada a los estudiantes y maestros vía Facebook para participar en el proyecto de localización.....	22
Figura 15. Actividades: a) inicio de sesión en la aplicación, b) selección de la ubicación, c) medición y recolección de datos, y d) desplazamiento.....	23
Figura 16. Estructura del árbol generado a partir de los datos obtenidos de uci data set Lenses.	30

Índice de Tablas

Tabla 1. Área y dimensiones de los escenarios de prueba.....	21
Tabla 2. Dispositivos móviles utilizados en la recolección de datos	25
Tabla 3. Muestras por salón	26
Tabla 4. Mediciones por dispositivo	26
Tabla 5. Pseudocódigo Clasificador k vecinos más cercanos	29
Tabla 6. Pseudocódigo de Árboles de Decisión ID3.....	31
Tabla 7. Paquetes de CRAN utilizados.....	33
Tabla 8. Valoración del índice Kappa (Fuente: Altman, 1991)	34
Tabla 9. Matriz de confusión para el modelo generado con un valor de k=3	35
Tabla 10. Matriz de confusión para el modelo generado con un valor de k=6	35
Tabla 11. Matriz de confusión para el modelo generado con un valor de k=9	36
Tabla 12. Matriz de confusión para el algoritmo J48	37
Tabla 13. Matriz de confusión para el algoritmo de Naive Bayes	38
Tabla 14. Rendimiento de los Algoritmos.....	39

Índice de Ecuaciones

Ecuación 1 Normalización de media cero	24
Ecuación 2. Ecuación para calcular la distancia Euclidiana	28
Ecuación 3. Ecuación para Entropía n clases	32
Ecuación 4. Entropía de los predictores	32
Ecuación 5. Ecuación para Ganancia de Información	32
Ecuación 6. Producto de funciones de probabilidad condicional	32
Ecuación 7. Probabilidad posteriori	33
Ecuación 8. Función para la asignación de una muestra x a una de las L clases existentes.....	33

Introducción

En este capítulo se describe la motivación de esta tesis y se introduce el tema de investigación objeto de ésta. Posteriormente, se plantean los objetivos de investigación, así como la metodología que se siguió durante esta tesis; y finalmente, se presenta la organización de este documento.

1.1 Motivación y contexto del problema

La localización se define como el espacio en el que algo o alguien se encuentra ubicado, en relación con otras cosas (Cambridge, 2017). Esta puede ser expresada en forma física utilizando el sistema de coordenadas; por ejemplo, la ciudad de Mexicali se encuentra a una latitud de 32°39.6' Norte y longitud de 115°28.5' Oeste. Otra forma de expresar la localización es a través del uso de etiquetas que describen el lugar; por ejemplo, sala, comedor, entre otros. A este tipo de localización se le conoce como simbólica (Hightower y Boriello, 2001).

La localización de un individuo es de gran interés, como información de contexto, para los desarrolladores de servicios basados en localización o sistemas conscientes de contexto (Dey et al., 2001). Esto se debe a la diversidad de aplicaciones que pueden desarrollarse, tales como, guías de turismo, sistemas de noticias, sistemas para asistir a personas, entre otros. Sin embargo, diseñar sistemas para conocer la ubicación de un individuo en un entorno de interiores, con base al Sistema Global de Posicionamiento (GPS) es un gran reto, ya que el desvanecimiento multitrayectoria, atenuación por distancia, o la interferencia provenientes de otros sistemas inalámbricos afectan la estimación de la localización (con errores de precisión de hasta 100 metros según el dispositivo)

(Bitar et al., 2006). Con un error de estimación de 50 metros, es probable que el individuo ya ni siquiera se encuentre ubicado en el interior del edificio o casa-habitación. Sin embargo, el desarrollo de nuevos y sofisticados sensores ha permitido que se propongan soluciones y estrategias que logran una mejor precisión de la localización de individuos en interiores. En estas se proponen utilizar dispositivos/sensores, tales como, *Wi-Fi*, *RFID*, *Bluetooth*, acelerómetros, magnetómetros, por mencionar algunos, así como diversas técnicas de posicionamiento, tales como, triangulación, proximidad o huella digital de señal, siendo esta última la más utilizada (Brena et al., 2017; Hightower et al., 2001).

Para obtener la huella digital de una señal se requieren dos fases: entrenamiento y cálculo de la posición. En la primera fase, se mide y recolecta la información asociada a una señal de interés en diferentes puntos del entorno de interiores (ver Figura 1). Esta información se almacena en una base de datos, conocida como *mapa de radio*. Para calcular la posición de un individuo, se mide y recolecta una señal de interés, ésta se compara con la información del mapa de radio, obtenido en la primera fase, utilizando un algoritmo de aprendizaje automático (p. ej. k vecinos más cercanos) (Taheri et al., 2004).

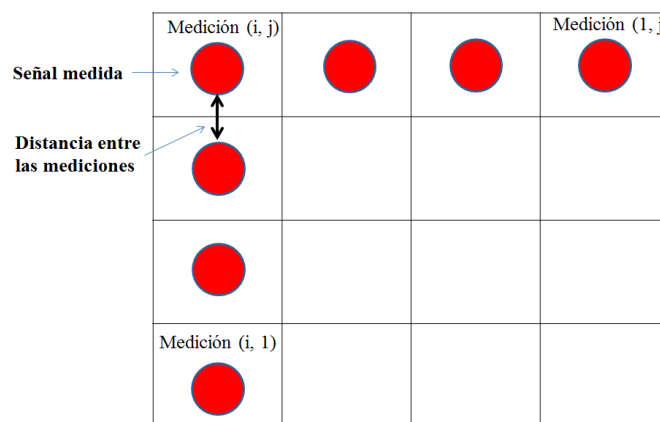


Figura 1. Entorno de interiores subdivido en regiones, en las cuales se mide y registra una señal considerando la distancia y orientación

Con la técnica de huella digital de señal es posible generar un modelo para estimar la localización de un individuo en un entorno de interiores (Taheri et al.,

2004, Galván-Tejada et al., 2014). Sin embargo, para aplicar esta técnica es necesario afrontar varios retos (Zheng et al., 2015; Shu et al. 2015, He et al., 2017), tales como: (i) se requiere que un experto realice la recolección de datos, ya que este tiene conocimientos sobre la señal de interés e instrumentos para medirla; (ii) en la etapa de entrenamiento, se requiere recolectar la mayor cantidad de posiciones posibles; y (iii) la dinámica del entorno afecta al mapa de radio generado; ya que cualquier cambio en la infraestructura del entorno de interiores requiere que la fase de recolección de datos se realice nuevamente.

Para abordar los retos que enfrentan los diseñadores de sistemas para localización en interiores al utilizar la técnica de huella digital de señal, se propone utilizar el paradigma de la colaboración abierta distribuida (Estellés-Arolas et al., 2012) como parte de la etapa de entrenamiento de la técnica.

1.2 Colaboración Abierta Distribuida

La colaboración abierta distribuida se define como un nuevo paradigma que empodera a los usuarios a contribuir en la realización de una tarea específica (Estellés-Arolas et al., 2012). Existen diversos proyectos que muestran el éxito de esta técnica en la tarea de recolección de información, por ejemplo, google crowdsource¹, OpenStreetMap² y Wikimapia³ (Ariffin, 2014). En estos proyectos los usuarios proporcionan información para enriquecer un mapa con notas o fotos, las cuales proporcionan información adicional del lugar (ver Figura 2).

¹ <https://crowdsourcing.google.com/>

² <https://www.openstreetmap.org/#map=5/23.944/-102.579>

³ <http://wikimapia.org/#lang=es&lat=32.651900&lon=-115.468300&z=12&m=b>

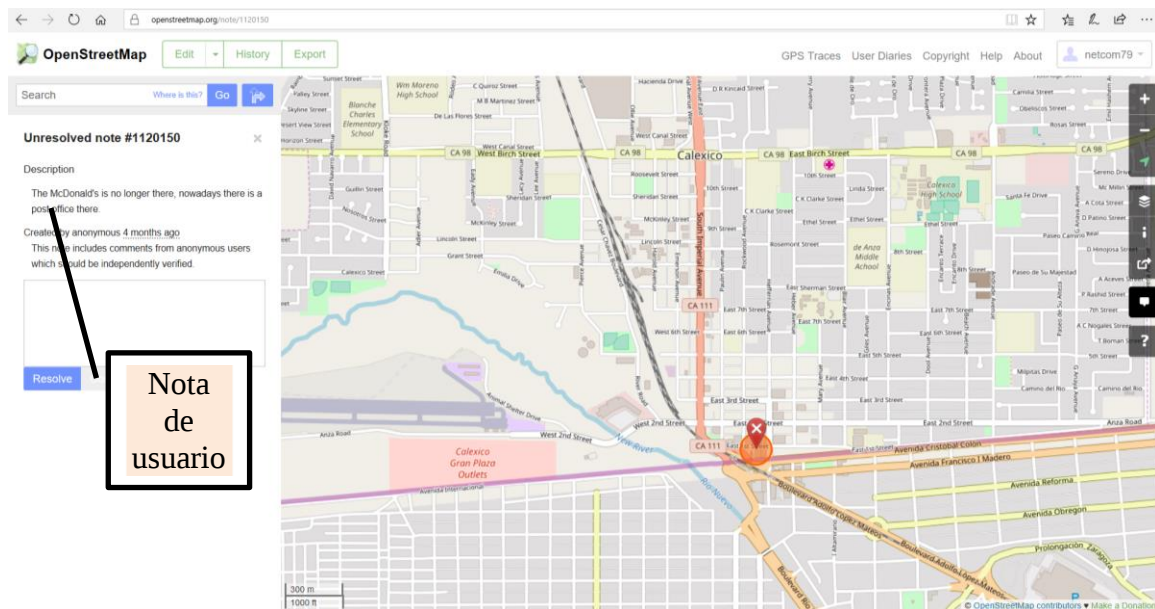


Figura 2. Nota de usuario que informa que el establecimiento comercial marcado en el mapa ya no existe

En el área de localización, la colaboración abierta distribuida, se define como el proceso en el cual varios individuos contribuyen con datos detectados o generados con sus dispositivos móviles, ya sean datos recolectados por medio de los sensores embebidos en el dispositivo móvil o fotografías (Rai et al., 2012; Luo et al., 2014). Información que posteriormente se utiliza para la construcción automática de planos virtuales de un entorno de interiores (Alzantot et al., 2012; Luo et al., 2014; Zhang et al. 2013); así como para la eliminación de la etapa de calibración en diseño de un sistema para localización en interiores (Yang et al., 2013; Rai et al., 2012).

En el presente trabajo de tesis, se propone utilizar la colaboración abierta distribuida como etapa de entrenamiento de la técnica de huella digital de señal, ya que de esta forma no se requiere que un experto realice la recolección de datos. Además, de que se podrá contar con la mayor cantidad de posiciones posibles con el fin de crear el mapa de radio y modelo predictivo para estimar la localización de un individuo en un entorno de interiores.

Como fuente de información, se propone utilizar el campo magnético de la Tierra, debido a que este se encuentra disponible en todos los entornos de interiores, y no requiere de una infraestructura adicional empotrada en el entorno, por ejemplo, puntos de acceso o radio bases; que estén generando una señal (Brena et al., 2017).

1.3 Objetivos de la Tesis

1.3.1 Objetivo General

El objetivo general de esta tesis es: *"Determinar la viabilidad de aplicar la colaboración abierta distribuida como etapa de recolección de datos en la técnica de huella digital de señal para el desarrollo de un sistema de localización para interiores"*.

1.3.2 Objetivos Específicos

Para alcanzar el objetivo general se proponen los siguientes objetivos específicos:

- Generar un mapa de radio con mediciones de intensidad de campo recolectadas mediante colaboración abierta distribuida.
- Crear un modelo predictivo para estimar la localización de un individuo en un entorno de interiores utilizando algoritmos de aprendizaje automático.
- Evaluar la eficiencia del modelo predictivo propuesto para estimar la localización de un individuo en un entorno de interiores.

1.4 Metodología

Para lograr los objetivos de investigación, se propone la siguiente metodología que consiste de cuatro etapas (ver Figura 3), la cual se basa en el modelo de referencia para la elaboración de proyectos de análisis de datos llamado *CRISP-DM* (del inglés, *Cross Industry Standard Process for Data Mining*) (Wirth y Hipp, 1999).



Figura 3. Metodología de Investigación

1.4.1 Recolección de datos

Para obtener el conjunto de datos experimental se desarrolló una aplicación en lenguaje JAVA versión 8.1, la cual utiliza el sensor magnetómetro empotrado en el dispositivo móvil para medir y recolectar la intensidad de campo magnético en μ Teslas en los ejes x, y, z.

La aplicación la utilizaron siete individuos quienes recolectaron datos en un entorno de interiores por un período de 10 días. Un total de 701,714 mediciones fueron recolectadas con 7 diferentes teléfonos inteligentes (Sony C6906, Sony D6503, Google Nexus 4, Google Nexus 7, Samsung SM-G531H, Verizon SM-G900V y Sony E5803). Los datos recolectados se separaron en dos subconjuntos de manera aleatoria. El primero sirve para entrenar a los algoritmos de aprendizaje automático y con esto proporcionen un modelo para localización en interiores.. Este está conformado por el 70% de los datos (491,200 mediciones). El segundo subconjunto, el cual comprende el 30% restante de los datos (210,514 mediciones) se utiliza para prueba y validación de los modelos predictivos.

1.4.2 Procesamiento de los datos

Para mejorar la calidad de los datos recolectados e incrementar la precisión de la estimación de la localización se realizaron ejecutaron dos procedimientos: 1). Los datos se procesaron mediante una herramienta de análisis de datos conocida como OpalConvert y 2) Los datos se normalizaron.

1.4.3 Generación del modelo de localización

En esta etapa se identificaron las relaciones entre las variables x, y, z y la norma para generar un modelo predictivo que ayude a estimar la localización de un individuo en un entorno de interiores. En total se generaron 5 modelos a partir de la información de intensidad de campo magnético recolectada. Tres de los modelos se generaron con el algoritmo de aprendizaje k Vecinos más Cercanos (Larose, 2005) (*k-NN* del inglés *k Nearest Neighbors*) utilizando los siguientes valores para k: 6, 3 y 9. Otro modelo se generó a partir de algoritmo de Árboles de Decisión (J48) (Rokach, 2005), y un quinto modelo probabilístico se generó a partir del algoritmo Bayesiano Ingenuo (*Naïve Bayes*) (Larrañaga, 1997). La implementación de los algoritmos se realizó a través del entorno de desarrollo integrado para R (RStudio).

1.4.4 Validación del modelo de localización

Para evaluar los modelos predictivos, se consideraron dos métricas que se obtienen a partir de la matriz de confusión, la exactitud global por clase y el coeficiente de concordancia (*kappa*). La exactitud global se refiere a la capacidad del modelo para predecir correctamente la localización de un individuo. Mientras que el coeficiente de *kappa* muestra la concordancia entre los datos de prueba y la clasificación hecha por el modelo (Rosado y Vergel, 2015). El modelo generado con el algoritmo de k Vecinos más Cercanos con un valor de k igual a 3 fue el que tuvo un mejor desempeño ya que ofrece una exactitud de 97.12 % y con un coeficiente de concordancia (*kappa*) de 0.9627 para estimar la localización de un individuo. El modelo predictivo con más bajo rendimiento fue el generado con el algoritmo de *Naïve Bayes*, ya que ofrece una exactitud de 50.79% con un nivel de concordancia de 0.3834 para estimar la localización de un individuo en interiores.

1.5 Organización de la Tesis

El presente manuscrito está organizado en 5 capítulos y 1 apéndice en los que se abordan los siguientes puntos:

Capítulo 2: presentan proyectos de investigación que utilizan el paradigma de colaboración abierta distribuida para capturar información en un entorno de interiores.

Capítulo 3: presenta la aplicación que se desarrollado para la captura de datos de intensidad de campo magnético. Así como se describe el proceso recolección de datos realizado y los resultados obtenidos.

Capítulo 4: se presentan los algoritmos de aprendizaje automático utilizados para la generación de los modelos predictivos. Así como la evaluación de los modelos generados considerando las métricas de exactitud y nivel de concordancia.

Capítulo 5: se presentan las conclusiones y las contribuciones de este trabajo de investigación. Así como se listan algunos nichos de oportunidad identificados para continuar con esta investigación.

Apéndice A: Datos de salida. Este presenta resultados de aspecto técnico y experimental que no fueron incluidos en el capítulo 5 para facilitar la lectura del documento.

Trabajo Relacionado

En este capítulo se describen algunos proyectos de investigación que han utilizado el paradigma de colaboración abierta distribuida en el área de localización para interiores; y se enfatizan las diferencias de este trabajo de tesis con dichas investigaciones previas.

2.1 Introducción

Se ha identificado que la colaboración abierta distribuida se utiliza en la localización objetos o personas en interiores como una herramienta para la recolección de información.

La colaboración de los individuos es básica, y esta recolección de datos se puede realizar bajo dos enfoques: (i) *consciente*, cuando el individuo participa aportando datos con su dispositivo móvil (Rai et al., 2012); o (ii) *inconsciente*, cuando los datos se recolectan de los dispositivos que porta un individuo sin que este tenga conocimiento de ello (Luo et al., 2014).

Los datos recolectados se utilizan para la construcción automática de planos virtuales de un entorno de interiores o para eliminar la etapa de calibración en el diseño de un sistema de localización. A continuación se describen estas actividades.

2.2 Sistemas para la construcción automática de planos

Estos sistemas se caracterizan por recolectar información referente al desplazamiento de un individuo en el interior de un espacio (p. ej. distancia, dirección o número de pasos). Así como información proveniente de sensores o

dispositivos que se encuentran embebidos en el entorno (p. ej. puntos de acceso). La información recolectada se utiliza para identificar las rutas que comúnmente emplean los individuos al desplazarse. Lo que permite proponer un plano virtual del entorno. A continuación se describen algunos ejemplos.

2.2.1 CrowdInside

En *Alzantot et al.* (2012) proponen el sistema llamado *CrowdInside*, el cual consiste de tres módulos: (a) módulo de recopilación de datos: los usuarios realizan la recolección consciente de mediciones de los sensores de acelerómetros, magnetómetros, giroscopios y los valores de intensidad de la señal WiFi recibidos desde puntos de acceso disponibles en el entorno. Este sistema utiliza GPS para la detección de la transición del usuario desde el exterior al interior; (b) En el módulo de generación de rastros se realiza la construcción de las trazas de movimiento basadas en una nueva técnica de restablecimiento de errores basada en anclajes, que son puntos en el entorno con firmas de sensor únicas. Estos fueron clasificados en dos clases: los basados en el sensor de GPS (entradas de edificios y ventanas) y los basados en sensores de inercia (escaleras, ascensores, escaleras mecánicas, puertas de habitaciones, etc.). Los autores comentan que debido a que el sistema es utilizado en el interior de un edificio, no puede depender del GPS ya que requiere de la visibilidad de los satélites. Por lo que proponen utilizar el enfoque basado en cálculos irreversibles, con la finalidad de identificar la transición del usuario desde el exterior al interior.

Se hace uso de otras técnicas como: la de clasificación, con el fin de separar los corredores de las salas y las técnicas de agrupación para la separación de las habitaciones. También se puede utilizar la técnica de geometría computacional que permite identificar las formas de las habitaciones, esta proporciona el diseño general de un edificio (ver Figura 4), en el que se identifican las ubicaciones, formas de los pasillos y otros puntos de interés, como ascensores, escaleras y escaleras mecánicas. Este sistema no requiere de una infraestructura especial

para la generación de trazos ni la estimación del plano, su implementación se realizó en teléfonos con sistema operativo Android; y (c) Módulo de Estimación del Plano de Piso: en este se realiza la separación de corredores de habitaciones y detecta los límites de las habitaciones. Para la obtención de los detalles de las partes internas del edificio, se utilizó un algoritmo de segmentación simple basado en el cambio de rumbo para segmentar las huellas y el filtrado de trazas para identificar los pasillos o habitaciones. También se utilizó un clasificador basado en la técnica de árbol estándar usando las características de tiempo promedio dedicado por paso en el segmento, longitud del segmento, densidad de rastreos de otros usuarios.

Para la evaluación, el sistema se implementó en diferentes teléfonos con sistema operativo Android (Samsung Nexus S, Nexus One, Galaxy Ace y Galaxy Tab). El experimento se llevó a cabo en dos ubicaciones: un centro comercial con muchas escaleras, ascensores, escaleras mecánicas y un edificio de un campus universitario con un área de aproximadamente 448m².

El primer banco de pruebas se usó para evaluar la precisión de la generación de trazas y el restablecimiento de errores basado en anclajes, obteniendo como resultados la detección de los puntos de anclaje con una tasa de FP de 0.2% y tasa de FN de 1.3%.

El segundo banco de pruebas se usó para evaluar el plano construcción.

Se recolectaron huellas con la ayuda de cuatro voluntarios que cubrían 12 habitaciones junto con todas las áreas de los pasillos. Como resultado se determinó que se requieren 290 segmentos para obtener la forma completa del plano.

Los autores concluyen que CrowdInside es único en la aplicación de los puntos de anclaje para mejorar la precisión de la recolección de huellas de señal.

Además, permite detectar automáticamente el contorno del plano y la forma detallada del plano.

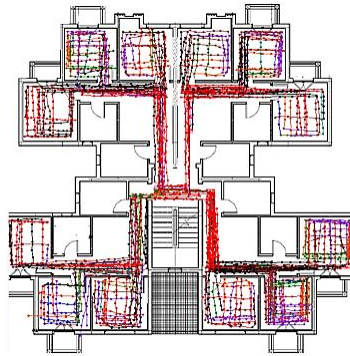


Figura 4. Rutas generadas con el sistema CrowdInside a partir del desplazamiento de un individuos dentro de un edificio (Fuente: Alzantot et al. 2012)

2.2.2 Piloc

En *Luo et al. (2014)* se describe el sistema Piloc, un sistema de localización para interiores que se basa en datos de velocidad angular y aceleración lineal generados de manera inconsciente por el usuario. PiLoc solo requiere de la disponibilidad de una infraestructura de red inalámbrica tipo WiFi.

Para la recolección de las trayectorias y muestras de la señal de WiFi, el usuario ejecuta un servicio en segundo plano en su dispositivo telefónico con sistema operativo Android. El sistema utiliza los datos recopilados para construir y actualizar los planos de planta periódicamente (ver Figura 5). Para cada consulta de localización, el servidor primero encuentra el mapa de radio correcto para usar en función del resultado de agrupamiento del punto de acceso. Posteriormente, se utiliza el algoritmo de coincidencia de trayectoria para encontrar la mejor ubicación de coincidencia de los datos enviados por dispositivo móvil.

El sistema Piloc se implementó y evaluó con datos recolectados a través de tres diferentes modelos de teléfonos inteligentes con sistema operativo Android(Google Galaxy Nexus, Samsung S3 y Samsung S4). Los autores

argumentan que el sistema puede posicionar a un sujeto dentro del entorno de interiores con un error de 1.5 metros en promedio.



**Figura 5. Rutas de un individuo asociadas a un punto de acceso
(Fuente: Luo et al., 2014)**

2.2.3 CIMLoc

CIMLoc es un sistema que utiliza información recolectada a través de la colaboración abierta distribuida para la construcción de un mapa digital (ver Figura. 6) (Zhang et al. 2013). Los datos recolectados por los individuos de manera consciente para la construcción del mapa provienen de la señal Wi-Fi, y de sensores tales como: magnetómetro, giroscopio y acelerómetro para la generación de las trayectorias que realizan los individuos de manera consciente con sus dispositivos móviles.

El teléfono inteligente obtiene un punto de referencia para relacionar las trayectorias e iniciar la construcción de un mapa. Los puntos de referencia son considerados a partir de la pérdida de la señal GPS. Para evaluar el sistema utilizaron dos teléfonos inteligentes con sistema operativo Android (HTC One X and Samsung Galaxy S4). Con estos recolectaron datos de las trayectorias recorridas por un individuo y se comparó con el mapa real del sitio digitalizado manualmente y el mapa digital construido por el sistema.

Los resultados de CIMLoc indican que este sistema es capaz de construir un mapa de alta apariencia al mapa real con un intervalo de confianza del 95%.

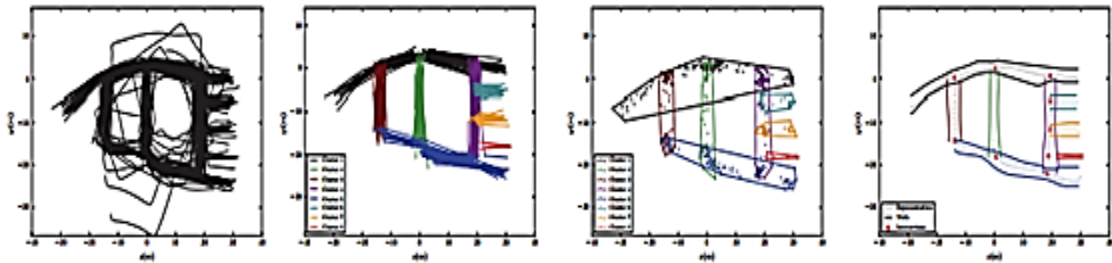


Figura 6. Mapa digital de un entorno de interiores construido con el sistema CIMloc (Fuente: Zhang et al. 2013)

2.2.4 Groping

En el trabajo de Zhang et al. (2015) se describe el sistema GROPING, su objetivo principal es el de construir de manera automática las rutas de un entorno de interiores mediante el uso de la colaboración abierta distribuida. Para el experimento se reclutaron 20 participantes, de los cuales 8 utilizaron sus dispositivos móviles para recolectar de manera consciente el desplazamiento de las diferentes rutas por donde pasaban, esto través del sensor magnetómetro de su dispositivo móvil (ver Figura 7). Cada mapa se enriquece con información semántica, es decir, marcas o etiquetas que describen el entorno, para facilitar la navegación. Este sistema se evaluó con datos recolectados de los 12 participantes restantes, los cuales utilizaron sus propios dispositivos móviles con sistema operativo Android de diferentes modelos (Samsung Galaxy S2 / S3 / Ace Plus, Sony Xperia S y HTC One X), demostrando la alta usabilidad de los servicios de navegación del sistema.

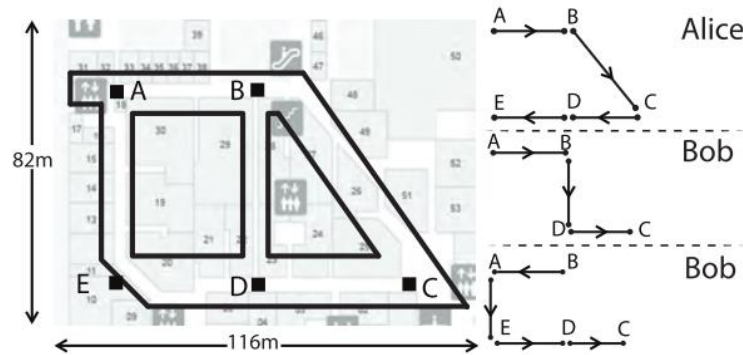


Figura 7. Construcción del plano de un edificio a partir de las trayectorias realizadas por de los individuos en el Proyecto GROPING (Fuente: Zhang et al. 2015)

2.3 Eliminación de la etapa de calibración

Para generar un modelo de localización en interiores de un sistema basado en radio frecuencia, es necesario realizar una fase previa de calibración del entorno de interiores. Se debe extraer un modelo estadístico de propagación de la señal, el cual se utiliza para derivar métodos de localización basados en comparación de patrones de radiación. Para reducir o evitar esta etapa de calibración se han utilizado los datos provenientes de la colaboración abierta distribuida. A continuación se presentan algunos proyectos.

2.3.1 Freeloc

En Yang et al. (2013) se describe un sistema de localización llamado *Freeloc*, el cual no requiere etapa de calibración, ya que utiliza la información recolectada de la intensidad de la señal de Wi-Fi a través de colaboración abierta para generar un modelo de propagación de la señal. Los autores utilizaron seis teléfonos inteligentes para recolectar firmas de señal de Wi-Fi en una ubicación particular (Ver Figura 8). Para la evaluación del sistema se recolectaron firmas de señal con cuatro diferentes teléfonos móviles en dos sitios del entorno. Los autores proponen su propio algoritmo que consiste en utilizar los valores de la fuerza de la señal que emiten los diferentes puntos de acceso en un edificio, se

realiza la comparación de la relación de la señal obtenida con la información de una ubicación en particular.

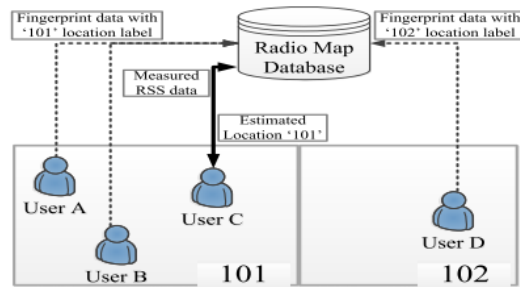


Figura 8. Funcionamiento del Sistema Freeloc (Fuente: Yang et al., 2013)

2.3.2 Zee

En el sistema Zee se recolectan datos de aceleración, dirección del desplazamiento de un individuo de manera inconsciente y de Wi-Fi para generar un mapa de radio que será utilizado para estimar la localización de un individuo (Rai et al., 2012). El objetivo de este sistema es recolectar los datos evitando la etapa de calibración mediante la caminata del usuario en un entorno de interiores. Para la recolección de los datos participaron seis individuos con sus dispositivos móviles para la recolección de los datos del acelerómetro y mediciones de la señal Wi-Fi (ver Figura 9). Para la evaluación del sistema Zee, se utilizó un dispositivo móvil para la recolección de los datos en el entorno de interiores. Sus resultados indican que es posible realizar localización con los datos provenientes de la colaboración abierta distribuida sin calibración.

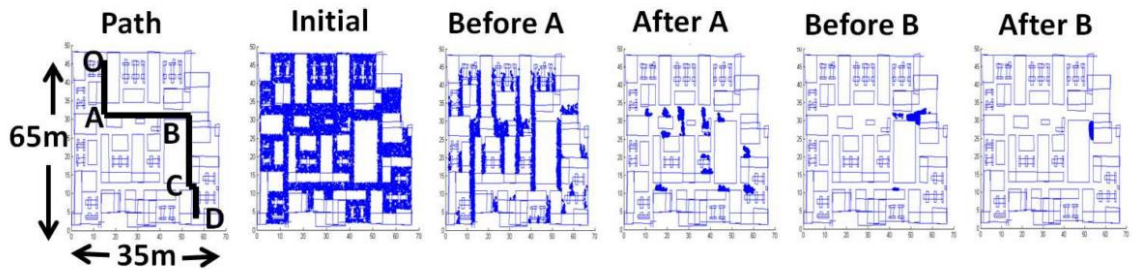


Figura 9. Ejemplo del Sistema Zee (Fuente: Rai et al., 2012)

2.3.3 MaWi

En Zhang et al. (2014) presentaron un sistema de localización en interiores llamado MaWi que se implementó para dispositivos móviles, este utiliza la fusión de la señal de Wi-Fi y el campo magnético de la tierra. Las firmas de señal se capturaron, por medio de una Colaboración Abierta Distribuida, empleando la secuencia de colección de muestra de datos magnéticos de una trayectoria en vez de un lugar como la firma de señal de Wi-Fi para eliminar la etapa de calibración. Para su evaluación se utilizaron dos dispositivos móviles con sistema operativo Android, concluyendo que se puede obtener una base de datos utilizable para la localización. Los autores comentan que una limitación de su trabajo es que la Colaboración Abierta Distribuida fue realizada por un solo usuario.

2.3.4 MaLoc

Este sistema se caracteriza por utilizar un enfoque sistemático para el diseño de un sistema de localización. Elimina la etapa de calibración de los sensores de diferentes dispositivos móviles para la recolección de la medición de campo magnético y la estimación de la longitud del paso del usuario (Xie et al., 2014).

Para su evaluación se realizó un experimento en el que participaron 7 personas con diferentes dispositivos móviles con sistema operativo Android (ver Figura 10).

A los participantes que se les colocaron estampas en los zapatos con la finalidad de poner un sello en el suelo durante su recorrido, y así evaluar los errores de conteo de pasos del sistema. Los autores demostraron que MaLoc es más preciso y eficiente que los sistemas basados en huellas digitales de Wi-Fi.

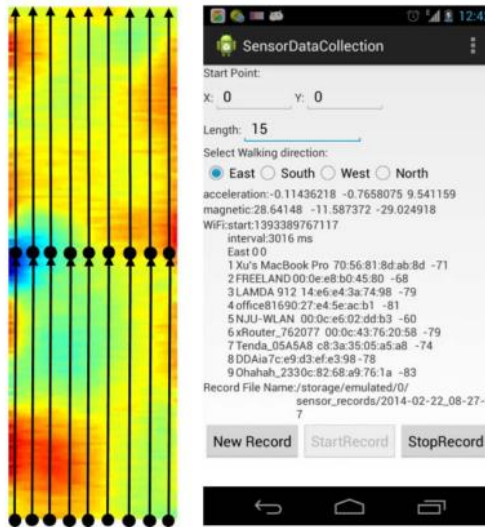


Figura 10. Interfaz del Sistema MaLoc

2.4 Conclusiones

En esta tesis a diferencia de los proyectos descritos en las secciones anteriores, se propone utilizar la colaboración abierta distribuida como etapa de entrenamiento de la técnica de huella digital de señal, con el fin de crear el mapa de radio y modelo predictivo para estimar la localización simbólica de un individuo en un entorno de interiores. Como fuente de información, se utilizará el campo magnético de la Tierra ya que se encuentra disponible en cualquier punto sobre la superficie terrestre, por lo que no se requiere infraestructura adicional (puntos de acceso o radio bases) para generarlo (Brena et al., 2017).

Recolección y procesamiento de datos

En este capítulo inicia con la presentación de la aplicación para dispositivos móviles que fue desarrollada para la capturar datos de intensidad de campo magnético. Enseguida, se describe el proceso que se realizó para la recolección de datos; y finalmente, concluye presentando resultados de la etapa de recolección de datos.-

3.1 Recolección de datos

Para recolectar los datos de intensidad de campo magnético en un entorno de interiores, se desarrolló una aplicación en lenguaje JAVA versión 8.0. La aplicación se ejecuta en dispositivos móviles inteligentes con sistema operativo Android versión 4.3 o superior. La aplicación utiliza el magnetómetro para medir y recolectar la intensidad de campo magnético en μ Teslas en tres diferentes ejes (x, y, z) (ver Figura 11); a una frecuencia de muestreo de 20 Hz.



Figura 11. Interfaz de usuario de la Aplicación

Adicionalmente, la aplicación también permite recolectar otros datos de interés,

tal como, la hora y fecha en que se realizó la medición, una etiqueta que indica el nombre del entorno de interiores en donde se realizó la medición (p. ej. salón A), el modelo de dispositivo móvil, la versión de sistema operativo, y el modelo del magnetómetro.

La información recolectada se almacena en un servicio en la nube llamado *Firebase*⁴. Este permite almacenar datos en un documento tipo *JSON* (del inglés: *JavaScript Object Notation*) con estructura en tipo de árbol, ver Figura 12 (Stonehemmm, 2016).

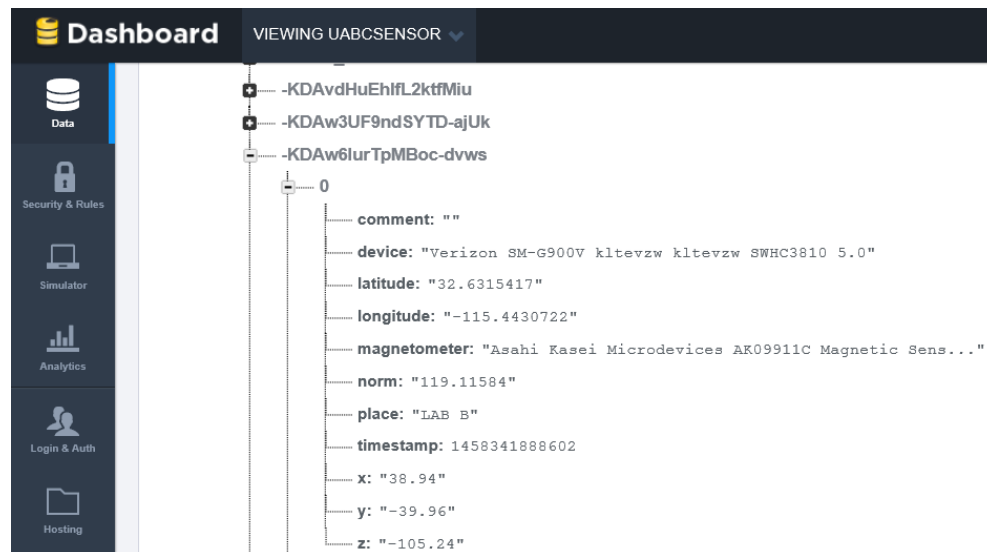


Figura 12. Estructura del documento *JSON* que almacena las mediciones

3.2 Descripción del entorno de experimentación

Como espacio de estudio se utilizó el Laboratorio de Sistemas Computacionales de la Facultad de Ingeniería de la UABC. El laboratorio se compone de salones y pasillos. Los salones considerados en el experimento se presentan en la Figura

⁴ <https://firebase.google.com/>

13, estos fueron etiquetados con las letras A, B, C, D y E. El área y dimensiones (largo y ancho) de los espacios se presentan en la Tabla 1.

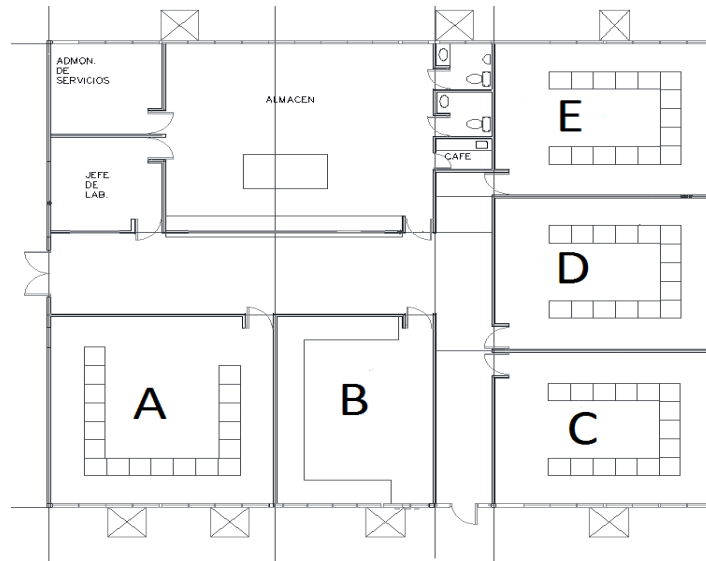


Figura 13. Plano del Laboratorio de Sistemas Computacionales

Tabla 1. Área y dimensiones de los escenarios de prueba

Salón	Dimensiones (Ancho x Largo)	Área
A	8m x 8m	64 m ²
B	4m x 8m	32 m ²
C	8m x 6.5m	55.25 m ²
D	8m x 6.5m	55.25 m ²
E	8m x 6.5m	55.25 m ²

3.3 Técnica para la recolección de datos

En este trabajo de tesis se aplicó la técnica de colaboración abierta distribuida para recolectar información del campo magnético. Se invitó a colaborar a estudiantes de la carrera de Licenciado en Sistemas Computacionales. La invitación se realizó vía una publicación en el muro de Facebook del programa

educativo de Licenciado en Sistemas Computacionales (ver Figura 14).

Para fomentar la permanencia y participación activa de los estudiantes en la tarea de recolección de datos, se les comentó a los participantes que se les invitaría a comer, y al participante que más muestras recolectará, se les recompensaría con dos boletos para el cine.

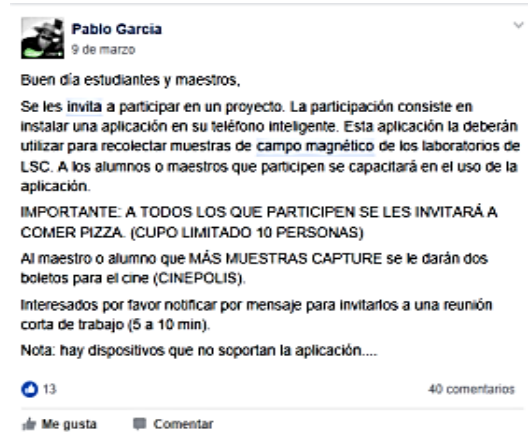


Figura 14. Invitación realizada a los estudiantes y maestros vía Facebook para participar en el proyecto de localización

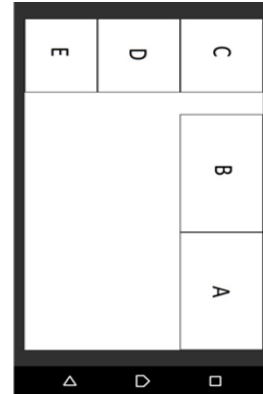
Para realizar la medición y recolección de la intensidad de campo magnético de un entorno de interiores; los alumnos fueron capacitados para utilizar la aplicación *Sensor.apk* la cual se encuentra disponible en el siguiente vínculo (<https://www.dropbox.com/s/vut1kprukhkn55I/Sensor.apk?dl=0>).

La capacitación de los estudiantes consistió de cuatro actividades: (1) iniciar la aplicación (ver Figura 15a), (2) indicarle como seleccionar en la aplicación el salón en el cual recolectarían la información (ver Figura 15b); (3) realizar la medición y recolección de información de campo magnético (ver Figura 15c); y finalmente (4) el desplazamiento en el entorno de interiores. Sobre este se les comentó a los estudiantes que deberían mantener su dispositivo móvil en la palma de su mano con la pantalla hacia arriba y desplazarse libremente por el salón (ver Figura 15d). Este proceso se repite cada vez que el alumno requiera

recolectar información.



a)



b)



c)



d)

Figura 15. Actividades: a) inicio de sesión en la aplicación, b) selección de la ubicación, c) medición y recolección de datos, y d) desplazamiento

3.4 Procesamiento de la información

Para asegurar la calidad de los datos capturados y lograr una mayor precisión en la estimación de la localización, se realizaron las siguientes actividades:

- **Formateo de los datos.** Los datos almacenados en el documento JSON se convierten a un formato tipo tabla (CSV, del inglés Comma Separate Values); en el que las columnas se separan por comas para poder

utilizarlo por herramientas para el análisis de información (p. ej. *RStudio*). Para realizar esto se utilizó la herramienta *Opal Convert*⁵.

- **Normalización.** Para asegurar que todos los atributos tengan la misma importancia, los datos fueron normalizados. El método utilizado fue normalización de media cero, ya que esta permite normalizar datos cuando se desconoce el valor mínimo y máximo de los atributos de interés (Han et al., 2011). En nuestro caso desconocíamos el mín./máx. de la intensidad de campo magnético en el edificio o en cada uno de los salones.

El cálculo de la normalización se realizó a partir de la Ecuación 1, donde v es un elemento del vector a normalizar, $\bar{A}$ y σ_A son la media y desviación estándar, respectivamente, y v' es un dato normalizado del atributo A .

$$v' = \frac{v - \bar{A}}{\sigma_A} \quad (1)$$

- **Partición de los datos.** Los datos recolectados se dividieron de manera aleatoria en dos subconjuntos de datos: entrenamiento y prueba. El primero se compone por el 70% de los datos y, se utiliza para entrenar los algoritmos de aprendizaje automático, con el propósito de generar un modelo predictivo de localización. Mientras que, el segundo subconjunto se generó con el resto de los datos (30%), y se utiliza para evaluar la exactitud del modelo predictivo.

3.5 Resultados

En total se recolectaron un total de 701,714 mediciones de intensidad de campo magnético utilizando 7 dispositivos móviles (ver Tabla 2).

⁵ <http://opal-convert-csv-to-qif-to-csv.en.softonic.com/>

Tabla 2. Dispositivos móviles utilizados en la recolección de datos

Dispositivo	Modelo de magnetómetro
Sony C6906	AK8963
Sony d6503	Ak09911
Google Nexus 4	LGE
Samsung SM-G531H	YAS537
Google Nexus 7	Invense MPL
Verizon SM_G900V	AK09911C
Sony E5803	Invense Inc.

En la Tabla 3 se presenta la cantidad de mediciones realizadas aleatoriamente en cada salón. En esta se puede observar que el salón con más mediciones fue el salón E, mientras que el salón con menos mediciones fue el A. Debido a que la recolección se realizó mediante la técnica de colaboración abierta distribuida, la aplicación desarrollada no obligaba a los participantes a recolectar el mismo número de muestras por salón.

Tabla 3. Muestras por salón

Salón	Número de mediciones
A	110,679
B	143,352
C	149,135
D	130,708
E	167,840

En la Tabla 4 se presentan las mediciones obtenidas por dispositivo móvil. En esta se puede observar que la mayoría de las muestras (34%) fueron recolectadas por el Sony modelo E5803.

Tabla 4. Mediciones por dispositivo

Dispositivo	Número de Mediciones
Sony C6906	13,700
Sony d6503	63,100
Google Nexus 4	50,500
Samsung SM-G531H	58,500
Google Nexus 7	77,000
Verizon SM_G900V	197,916
Sony E5803	240,998

Generación del Modelo para Localización

En este capítulo se presentan los modelos de localización para interiores que fueron generados a partir de la información de campo magnético recolectada a través de la técnica de colaboración abierta distribuida.

4.1. Introducción

En el contexto del diseño de sistemas de localización para interiores, los algoritmos de aprendizaje automático han sido utilizados como herramienta para la generación de modelos predictivos, los cuales tienen por objetivo estimar la localización de un individuo en un entorno en interiores considerando un mapa de radio. En esta tesis se utilizaron los siguientes algoritmos de aprendizaje automático para generar modelos para localización en interiores: k Vecinos más Cercanos (k-NN), Árboles de Decisión (J48), y Bayesiano Ingenuo (Naïve Bayes). Estos fueron seleccionados debido a que en la literatura se reporta que forman parte de los diez algoritmos más utilizados para la generación de modelos de clasificación (Wu et al., 2007); y además, porque en diversos trabajos del área de localización en interiores se reporta que tienen un buen desempeño para generar modelos de localización basados en campo magnético (Galván-Tejada et al., 2014, Chung et al., 2011, Storms et al., 2010). En las siguientes secciones se describe brevemente cada uno de los algoritmos utilizados, y se establecen los parámetros que fueron utilizados en los experimentos.

4.1.1 Clasificador K vecinos más cercanos

El algoritmo de K vecinos más cercanos almacena todos los casos disponibles y clasifica nuevos casos con base a una medida de similitud (p. ej. medida de

distancia) (Witten y Frank, 2005). En este trabajo de tesis se consideró utilizar la medida de distancia Euclidiana, la cual se puede calcular empleando la Ecuación 2. En general la distancia euclidiana entre dos puntos $P = (p_1, p_2, \dots, p_n)$ y $Q(q_1, q_2, \dots, q_n)$, del espacio euclídeo n-dimensional, se define por;

$$D_E(P, Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_1^n (p_i - q_i)^2} \quad (2)$$

Para seleccionar el valor el óptimo de k se consideraron las heurísticas propuestas por Kirk (2015):

- (1) Evitar un número para cuando solo existen dos clases;
- (2) Seleccionar un valor de k que es más grande o igual al número de clases más 1; y
- (3) Seleccionar una k pequeña para evitar el ruido.

El pseudocódigo para implementar este algoritmo es presentado en la Tabla 5. En esta se indica que se tiene una entrada de datos, la cual consta de atributos asociados a una clase $D = \{(x_1, c_1), \dots, (x_N, c_N)\}$. Así como un nuevo dato que se desea clasificar x , el cual consta de atributos $x = (x_1, \dots, x_n)$ pero no cuenta con el valor de la clase. Para asociarlo a una clase, se sugiere como primer paso calcular la distancia entre los atributo del nuevo registro y los atributos de cada registro de la entra de datos $d_i = d(x_i, x)$. Posteriormente, se ordenan las distancias y se eligen varios valores considerando el valor de k. La clase que aparece con más frecuencia es la asignada al nuevo registro D_x^K .

Tabla 5. Pseudocódigo Clasificador k vecinos más cercanos

COMIENZO

Entrada: $D = \{(x_1, c_1), \dots, (x_N, c_N)\}$

$x = (x_1, \dots, x_n)$ nuevo caso a clasificar

PARA todo objeto ya clasificado (x_i, c_i)

calcular $d_i = d_E(x_i, x)$

Ordenar $d_i (i = 1, \dots, N)$ en orden ascendente

Dejar los K casos D_x^K ya clasificados más cercanos a x

Asignar a x la clase más frecuente D_x^K

FIN

Fuente (Stanton, 2013).

4.1.2 Árboles de decisión

Árboles de decisión incorporan un enfoque de clasificación supervisada (Witten y Frank, 2005). La idea procede de la estructura ordinaria del árbol que se compone de una raíz y nodos internos (posiciones en donde se dividen las ramas), las ramas y las hojas. De manera similar, un árbol de decisión se construye a partir de nodos que representan círculos y las ramas están representadas por los segmentos que conectan los nodos. Un árbol de decisión empieza desde la raíz, se mueve hacia abajo y generalmente se dibuja de izquierda a derecha. El nodo de donde empieza el árbol se llama nodo raíz. El nodo donde termina la cadena se conoce como nodo "hoja". Dos o más ramas pueden extenderse desde cada nodo interno, es decir, un nodo que no es nodo de hoja. Un nodo representa una cierta característica mientras que las ramas representan un rango de valores. Estos rangos de valores actúan como puntos de partición para el conjunto de valores de la característica dada. La Figura 16

muestra la estructura del árbol de conjunto de datos llamado Lenses⁶. Este árbol de decisión permite resolver sobre el tipo de lente de contacto mas adecuado para un paciente.

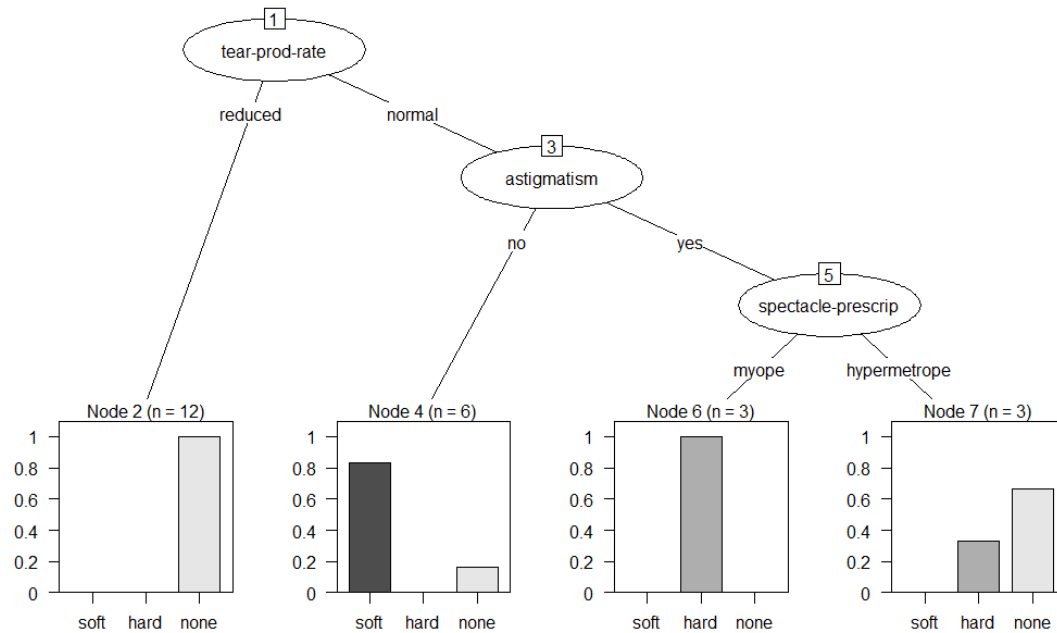


Figura 16. Estructura del árbol generado a partir de los datos obtenidos de uci data set Lenses.

El agrupamiento de datos en el árbol de decisión se basa en los valores de los atributos de los datos dados. Un árbol de decisión se forma a partir de los datos pre-clasificados. La división en clases se decide sobre las características que mejor divide los datos. Los elementos de datos se dividen según los valores de estas características. Este proceso se aplica a cada subconjunto dividido de los elementos de datos recursivamente. El proceso termina de la misma forma que todos los elementos de datos del subconjunto actual pertenecen a la misma clase.

En este trabajo de tesis se consideró utilizar el algoritmo ID3. Este algoritmo emplea una búsqueda descendente.

⁶ <https://archive.ics.uci.edu/ml/datasets/Lenses>

Tabla 6. Pseudocódigo de Árboles de Decisión ID3

ID3 (Conjunto de datos aprendizaje S, Conjunto de atributos A, Valores de atributos V)

Retorno Arbol de decisión.

Comienzo

Cargar primero conjuntos datos aprendizaje (S),
crear la raíz del árbol de decisión
nodo 'nodo raíz', agregue el conjunto de aprendizaje S al nodo raíz
como su subconjunto.

Para el nodo raíz, calcular
Entropía(subconjunto.nodoraíz) primero

Si Entropía(subconjunto.nodoraíz)=0, entonces
subconjunto.nodoraíz consiste en registros todos con el mismo
valor para el atributo categórico, devolver una hoja nodo con
decisión **atributo:atributo valor;**

Si Entropía (subconjunto.nodoraíz) != 0, entonces
Calcular ganancia de información para cada
atributo izquierdo (no utilizados en la division)
encontrar atributo A con Máxima(Ganacia(S,A))

Crear nodo hijo del nodo raíz y agregar a nodoraíz en el
arbol de decision.

Para cada hijo de nodoraíz, aplicar
ID3(S,A,V) recursividad hasta alcanzar
Nodo con entropía =0 o alcanzar nodo hoja.

Fin ID3.

Fuente: (Peng et al., 2009)

El pseudocódigo para implementar este algoritmo es presentado en la Tabla 6. En esta se indica que como entrada se tienen al conjuntos datos de aprendizaje (S) para crear el nodo raíz, y el primer paso es calcular la entropía con las formulas (3) y (4); la primera para calcular la entropía de las clases y la segunda para calcular la entropía de cada atributo con respecto a las clases. El segundo paso, consiste en encontrar la ganancia de la información de cada uno de los atributos, para calcular esto se utiliza la ecuación (5).

$$\text{Entropía (S)} = - \sum_{i=1}^c (P_i) \log_2 (P_i) \quad (3)$$

$$\text{Entropía(S, V)} = \sum_{c \in (V)} P(A) * \text{Entropía (V)} \quad (4)$$

$$\text{Ganancia (S, A)} = \text{Entropía(S)} - E(S, V) \quad (5)$$

En donde (S) equivale al conjunto de datos de aprendizaje, (A) al conjunto de atributos, (V) a los valores de cada atributo, (c) Clases y P_i es probabilidad de ocurrencia.

4.1.3. Naïve Bayes

Este algoritmo de aprendizaje automático está construido sobre modelos probabilísticos y, en concreto, hace uso del teorema de Bayes sobre las probabilidades condicionadas.

Este clasificado asume, para una muestra x , que sus atributos $x_1, x_2, \dots, x_n$ presentan una independencia condicional dado el valor de la clase, por lo que la probabilidad condicional puede expresarse como el producto de funciones de probabilidad condicional de cada atributo por separado;

$$p(x|\omega_i) = \prod_{j=1}^n p(x_j|\omega_i) \quad (6)$$

Usando el teorema de Bayes, la probabilidad posteriori se escribe como

$$p(x|\omega_i) = p(\omega_i) \prod_{j=1}^n p(x_j|\omega_i) \quad (7)$$

Finalmente, el algoritmo de Naive Bayes asigna una muestra x a una de las L clases existentes usando la siguiente función;

$$\omega^* = \operatorname{argmax}_{\omega_j} p(\omega_j) \prod_{j=1}^n p(x_j|\omega_j) \quad (8)$$

4.2 Implementación de los Algoritmos de Aprendizaje Automático

La implementación de los algoritmos de aprendizaje automático se realizó a través del entorno de desarrollo integrado para R (RStudio versión 1.0.136). Los paquetes utilizados se listan en la Tabla 7.

Tabla 7. Paquetes de CRAN utilizados

Algoritmo	Paquete
k-NN	<i>class</i> versión 7.3-14
Naïve Bayes,	<i>e1071</i> versión 1.6-7
Árboles de decisión (J48)	<i>RWeka</i> versión 0.4-26, <i>RWekajars</i> versión 3.7.13-1, <i>rpart</i> 4.1-10 y <i>party</i> versión 1.0-25 y <i>partykit</i> versión 1.0-5

4.3 Validación del mapa de radio

Para evaluar el desempeño de los algoritmos de aprendizaje, en esta tesis se consideraron las siguientes métricas calculadas a partir de la matriz de confusión: exactitud y coeficiente de Kappa. La exactitud se refiere a la capacidad del modelo para predecir correctamente la localización de un individuo. Mientras que el coeficiente de Kappa muestra la concordancia entre los datos de prueba y la clasificación hecha por el modelo (Rosado y Vergel, 2015).

Para interpretar el valor de Kappa se consideró la escala propuesta por Altman (1991) que se presenta en la Tabla 8.

Tabla 8. Valoración del índice Kappa (Fuente: Altman, 1991)

Valor de Kappa	Fuerza de concordancia
< 0.20	Pobre
0.21 – 0.40	Débil
0.41 – 0.60	Moderada
0.60 – 0.80	Buena
0.81 – 1.00	Muy Buena

4.4 Resultados

Un total de cinco modelos fueron generados utilizando los datos de entrenamiento. Se generaron 3 modelos con el algoritmo de k-NN con los valores de k 3,6 y 9.

Los modelos fueron evaluados con 210,512 mediciones (30% de los datos recolectados). A continuación se presentan las matrices de confusión de cada modelo.

Tabla 9. Matriz de confusión para el modelo generado con un valor de k=3

Referencia								
Predicción	LAB A	LAB B	LAB C	LAB D	LAB E	Total	%CORRECTAS	%INCORRECTAS
LAB A	37,688	288	608	349	366	39,299	95.90%	4.09%
LAB B	281	32,512	166	177	84	33,220	97.86%	2.13%
LAB C	552	169	48,961	359	219	50,260	97.41%	2.58%
LAB D	353	149	417	43,402	447	44,768	96.94%	3.05%
LAB E	338	85	200	453	41,889	42,965	97.49%	2.30%
	TOTAL					210,512	97.12%	2.83%

En la matriz de confusión se observa que de los 39,299 de datos de la clase LAB A, 37,688 datos fueron clasificados correctamente con un 95.90% y 4.09% de manera incorrecta (ver Tabla 9.)

Tabla 10. Matriz de confusión para el modelo generado con un valor de k=6

Referencia								
Predicción	LAB A	LAB B	LAB C	LAB D	LAB E	TOTAL	%CORRECTAS	%INCORRECTAS
LAB A	37,345	210	736	288	319	38,898	96.01%	4.0%
LAB B	306	32,886	193	228	115	33,728	97.50%	2.50%
LAB C	529	122	48,599	311	193	49,754	97.70%	2.33%
LAB D	427	103	503	43,471	427	44,931	96.80%	3.20%
LAB E	370	75	258	535	41,965	43,203	97.13%	2.69%
	TOTAL					210,512	97.03	2.96%

En la matriz de confusión se observa que de los 38,898 de datos de la clase LAB A, 37,345 datos fueron clasificados correctamente en un 96.01% y 4.09% de manera incorrecta (ver Tabla 10.)

Tabla 11. Matriz de confusión para el modelo generado con un valor de k=9

Referencia								
Predicción	LAB A	LAB B	LAB C	LAB D	LAB E	TOTAL	%CORRECTAS	%INCORRECTAS
LAB A	37,507	284	693	415	389	39,288	95.46%	4.53%
LAB B	266	32,531	147	170	95	33,209	97.95%	2.04%
LAB C	725	180	48,791	397	265	50,358	96.88%	3.11%
LAB D	386	127	476	43,287	485	44,761	97.70%	3.29%
LAB E	328	81	245	471	41,771	42,896	97.37%	2.43%
					TOTAL	210,512	96.88%	3.11%

En la matriz de confusión se observa que de los 39,288 de datos de la clase LAB A, 37,507 datos fueron clasificados correctamente en un 95.46% y 4.53% de manera incorrecta (ver Tabla 11.)

Considerando los valores de cada matriz de confusión se obtuvieran las siguientes métricas. La exactitud del modelo con k=3 obtenidos para este algoritmo k-NN fue de 97.12% con una kappa de .9639, para k=6, 97.03% con una kappa de .9627 para k=9 y 96.88% con una kappa de .9189. Se puede observar que al utilizar k=3 se obtiene mejor precisión ya que entre menor sea el valor de k será menor el procesamiento del algoritmo.

Para el algoritmo J48 se construyó un árbol de decisión a partir de las 701,714 instancias recolectadas (70% para entrenamiento y 30% para evaluación).

El proceso de prueba arrojó que el 96.91 % de las instancias fueron clasificadas correctamente, mientras que el 3.09% se clasifican de manera incorrecta.

Tabla 12. Matriz de confusión para el algoritmo J48

Referencia								
Predicción	LAB A	LAB B	LAB C	LAB D	LAB E	TOTAL	%CORRECTAS	%INCORRECTAS
LAB A	37,498	300	666	406	342	39,212	95.63%	4.37%
LAB B	258	32,564	150	156	75	33,203	98.08%	1.92%
LAB C	709	169	48,779	435	260	50,352	96.88%	3.12%
LAB D	332	175	435	43,319	479	44,740	96.82%	3.27%
LAB E	378	91	263	497	41,776	43,005	97.14%	2.86%
	TOTAL					210,512	96.91%	3.09%

En la matriz de confusión se observa que de los 38,898 de datos de la clase LAB A, 37,345 datos fueron clasificados correctamente en un 96.01% y 4.09% de manera incorrecta (ver Tabla 12.)

En el caso del bayesiano ingenuo considerando los valores de cada matriz de confusión se obtuvieran las siguientes métricas. La exactitud del modelo obtenido para este algoritmo Naïve Bayes fue: exactitud de 50.77% con una Kappa de 0.3838.

Tabla 13. Matriz de confusión para el algoritmo de Naive Bayes

Referencia								
Predicción	LAB A	LAB B	LAB C	LAB D	LAB E	TOTAL	%CORRECTAS	%INCORRECTAS
LAB A	12,688	7,278	4,092	7,563	7,591	39,212	32.35%	67.64%
LAB B	4,602	1,778	3,172	3,199	4,452	33,203	53.54%	46.45%
LAB C	5,418	4,799	20,179	14,126	5,830	50,352	40.07%	59.95%
LAB D	5,192	3,036	5,030	28,265	3,217	44,740	63.17%	36.82%
LAB E	6,279	2,016	3,913	2,954	27,843	43,005	64.74%	30.56%
	TOTAL					210,512	50.77%	48.28%

En la matriz de confusión e observa que de los 39,212 de datos de la clase LAB A, 12,688 datos fueron clasificados correctamente en un 32.35% y 67.64% de manera incorrecta (ver Tabla 13.)

Un resumen de los resultados obtenidos con las distintas técnicas de clasificación se presentan en la tabla 14.

Tabla 14. Rendimiento de los Algoritmos

Algoritmo	Instancias correctamente clasificadas en %	Instancias correctamente clasificadas en %	Coeficiente Kappa en %
KNN (k=6)	97.03 %	2.96 %	96.27 %
KNN (k=3)	97.12 %	2.88 %	96.39 %
KNN (k=9)	96.88 %	3.11 %	91.89 %
J48	96.91 %	3.09 %	96.17 %
Naïve Bayes	50.79 %	49.21 %	38.34 %

Los modelos con mejor rendimiento se obtuvieron a partir de los algoritmos de KNN con k=3 y J48, ya que ofrecen una exactitud de 97.12 % y 96.91 % respectivamente para localizar a un individuo. El modelo con el rendimiento más bajo se obtuvo con el algoritmo bayesiano ingenuo (Naïve Bayes), ya que ofrece una exactitud del 50.79% y un nivel de concordancia de 0.3834.

Debido a que los atributos se tratan como si fueran independientes dada la clase, la adición de datos redundantes sesga el proceso de aprendizaje (Witten y Frank, 2005), como un ejemplo extremo, si incluyese un nuevo atributo con los mismos valores de temperatura que los datos meteorológicos, el efecto del atributo temperatura se multiplicaría: todas sus probabilidades se cuadrarían, dado que es una influencia mucho mayor en la decisión.

Conclusiones y Trabajo Futuro

5.1. Conclusiones

En este trabajo de investigación se describe un estudio para determinar la viabilidad de la colaboración abierta distribuida como etapa de recolección de datos en la técnica de huella digital de señal. Como parte del experimento se entrenaron diversos algoritmos de aprendizaje supervisado para generar un modelo predictivo para la localización en interiores utilizando datos de intensidad de campo magnético recolectados en diversos salones de un edificio.

Los resultados obtenidos demuestran que es posible generar un mapa de radio y modelo predictivo con datos de intensidad de campo magnético recolectados a través de la técnica de colaboración abierta distribuida. Como resultado se obtuvieron cinco modelos. El que ofrece una mayor exactitud para localizar a un individuo en interiores fue el modelo obtenido con el algoritmo de k vecinos más cercanos (k -NN) con $k=3$, ya que ofrece una exactitud del 97.12% con un nivel alto de concordancia, $\kappa=.9639$. El modelo con el más bajo rendimiento fue el que se obtuvo con el algoritmo Bayesiano ingenuo (Naïve Bayes), ya que ofrece una exactitud del 50.79% con un bajo nivel de concordancia, $\kappa=.3834$. Por lo que el modelo generado no es útil para estimar la localización de un individuo en un entorno de interiores.

A continuación en la siguiente sección se presentan las contribuciones obtenidas durante este trabajo de tesis, así como los posibles nichos de oportunidad para continuar con este trabajo de investigación.

5.2. Contribuciones

Las contribuciones obtenidas como resultado de esta investigación, se presentan en las siguientes subsecciones:

5.2.1 Publicaciones de la Tesis

Con los resultados de la tesis se obtuvieron los siguientes artículos:

- Isabel L. Ambriz Silva, Juan P. García Vázquez, Ángel G. Andrade (2018). **Localización en Interiores vía Colaboración Abierta Distribuida y Campo Magnético**, por definir

En este artículo de revista se presenta un comparativo del rendimiento entre 5 modelos predictivos para la localización de un individuo en interiores, los cuales fueron creados a partir de los datos de intensidad de campo magnético recolectados a través de colaboración abierta distribuida. Para crear los modelos se utilizaron diversos algoritmos de aprendizaje automático: k-vecinos más próximos (con valores de $k=3, 6$ y 9), árboles de decisión J48 y bayesiano ingenuo. Los algoritmos fueron implementados en el entorno de desarrollo integrado para R (RStudio). Para evaluar el rendimiento de los algoritmos se consideraron las métricas de exactitud y el coeficiente Kappa.

- Isabel L. Ambriz Silva, Juan P. García Vázquez, Ángel G. Andrade, Luis Vizcarra Corral. (2016). **Algoritmo de Clasificación para Generar un Mapa de Radio para Localización en Interiores**, Presentado en el Encuentro Nacional de Ciencias de la Computación (**ENC 2016**), Taller de Inteligencia Artificial: Métodos y Aplicaciones, del 14 al 16 de Noviembre de 2016 en Chihuahua, Chihuahua, ISBN:978-607-9424-94-7.

En este artículo de congreso se describe un experimento en el que creo un modelo predictivo para la localización de un individuo en interiores utilizando los datos de intensidad de campo magnético recolectados a través de colaboración abierta distribuida. Para crear el modelo se utilizó el algoritmo de k vecinos más cercanos (k-NN), el cual fue implementado en el entorno de desarrollo integrado

para R (RStudio). Para evaluar el rendimiento del algoritmo se consideró la métrica de exactitud.

- Isabel L. Ambriz Silva, Juan P. García-Vázquez, Luis Vizcarra-Corral. (2016). **Generando un mapa de radio para localización en interiores mediante la colaboración abierta distribuida**, Presentado en el Congreso Internacional en Ciencias Computacionales (**CICOMP 2016**), del 23 al 28 de Octubre en Ensenada B.C, 2016, 346-349, ISBN-13: 978-1540303967.

En este cartel se presenta un modelo predictivo para la localización de un individuo en interiores utilizando los datos de intensidad de campo magnético recolectados a través de colaboración abierta distribuida. Los datos fueron divididos en entrenamiento y prueba utilizando el método de validación cruzada. Para crear el modelo predictivo se utilizó el algoritmo de árboles de decisión J48, el cual fue implementado en el entorno para análisis del conocimiento (Weka). Para evaluar el rendimiento del algoritmo se consideró la métrica de exactitud.

- Isabel L. Ambriz Silva. **Sensado participativo para desarrollar un sistema de localización para interiores que utiliza el campo magnético de la tierra**, Presentado en el marco de la 17 Feria de Posgrado de Calidad, del 22 al 23 de abril del 2016 en Ciudad de México.

En este cartel se presenta el protocolo de investigación, y la aplicación desarrollada para la recolección de datos. Se presenta como resultado la recolección realizada por seis estudiantes, obteniendo un total de 459,996 muestras de campo magnético con diferentes dispositivos.

5.3 Trabajo Futuro

5.3.1 Modelar la señal de campo magnético a través de características temporales, de frecuencia y estadísticas

Durante el procesamiento de los datos, se identificó que existe una gran variación entre las mediciones de intensidad de campo magnético recolectadas entre los magnetómetros de los distintos dispositivos móviles. Por lo que se tiene la hipótesis de que el rendimiento del algoritmo disminuirá cuando se trate de localizar a un dispositivo móvil que no esté registrado en el conjunto de datos de entrenamiento, ya que las muestras recolectadas por magnetómetro podrían estar fuera del rango de los valores de intensidad registrados en el mapa de radio. Considerando lo anterior, se propone como trabajo futuro trabajar en el procesamiento de la señal de intensidad de campo magnético para extraer y seleccionar características temporales, de frecuencia y estadísticas que permitan modelar la señal de intensidad de campo magnético. Esto a través del uso de ondoletas (Nerguizian et al., 2007) y/o transformada rápida de Fourier (Zhang et al., 2006).

5.3.2 Generar un modelo predictivo para entornos de interiores multi-nivel.

Durante la recolección de datos se identificó que el valor del eje Z del magnetómetro depende de la altura a la que se encuentra el dispositivo. Por lo que podría servir para identificar el nivel en el que se encuentra el sujeto.

5.3.3 Tamaño de los salones

Durante la experimentación de los diferentes algoritmos, se observó que el total de las muestras variaba independientemente al tamaño del salón. Por lo que se propone establecer la cantidad de muestras necesarias para general el modelo de predicción con base a las dimensiones de los salones.

5.3.4 Identificar las variaciones del campo magnético en el entorno de interiores por objetos presentes en el entorno.

Durante la recolección de los datos se observó que los parámetros X y Y del campo magnético varían considerablemente al estar presentes objetos que contienen metal tales como; computadoras o centros de carga. Por lo que se propone evaluar los modelos de predicción considerando la presencia de objetos para identificar su influencia sobre las variaciones del campo magnético de los objetos.

Referencias Bibliográficas

- Ariffin, B. Solemon and W. M. L. Wan Abu Bakar, (2014). An evaluative study on mobile crowdsourcing applications for crime watch, *In Proceedings of the 6th International Conference on Information Technology and Multimedia*, Putrajaya, pp. 335-340.
doi: 10.1109/ICIMU.2014.7066655
- Alzantot M., Youssef M. (2012) Crowdinside: automatic construction of indoor floorplans. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pp. 99–108.
- Bitar N. A., Escher A. C., Macabiau C., Boucheret M. L. (2006). Analysis of the gps acquisition environment: indoors and outdoors. In *Proceedings of 2nd International. Conference on Information Comm. Tech.*, 2006(1): 127–132.
- Brena R., García-Vázquez J. P., Galván-Tejada C. E., Muñoz-Rodríguez, D., Vargas-Rosales C., Fangmeyer J. (2017). Evolution of Indoor Positioning Technologies: A Survey. *Journal of Sensors*, Hindawi, pp. 1-21.
- Chen W., Guan M., Wang L., Ruby R., Wu K. (2017). FLoc: Device-free passive indoor localization in complex environments, *2017 IEEE International Conference on Communications (ICC)*, Paris, 2017, pp. 1-6.
doi: 10.1109/ICC.2017.7997098
- Dey A. K., Abowd G. D., Salber D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-computer interaction*, 16(2):97–166.
- Jaewoo Chung, Matt Donahoe, Chris Schmandt, Ig-Jae Kim, Pedram Razavai, and Micaela Wiseman. (2011). Indoor location sensing using geo-magnetism. In *Proceedings of the 9th international conference on Mobile systems, applications, and services (MobiSys '11)*. ACM, New York, NY, USA, 141-154. DOI=<http://dx.doi.org/10.1145/1999995.2000010>
- Diccionario Cambridge vincula
<http://dictionary.cambridge.org/dictionary/english/position> accedido el 07 agosto de 2017: 4:23 p.m.
- Estellés-Arolas, E., González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2), 189-200.

- Faria, J., Lopes, S., Fernandes, H., Martins, P., & Barroso, J. (2010, September). Electronic white cane for blind people navigation assistance. In *World Automation Congress (WAC)*, IEEE , 2010,pp. 1-7..
- Galván-Tejada, C. E., García-Vázquez, J. P., Brena, R. (2014) Magnetic field feature extraction and selection for indoor location estimation. *Journal of Sensors*, 14(6):11001-11015.
- Guan R. and R. Harle, "Towards a crowdsourced radio map for indoor positioning system," In *Proceedings. of 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, Kona, HI, 2017, pp. 207-212. doi: 10.1109/PERCOMW.2017.7917559.
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier, pág 71.
- Hightower J., Borriello, G. (2001). Location systems for ubiquitous computing. *Computer*, 172 34(8):57–66.
- He S., Shin K. G. (2017). Geomagnetism for Smartphone-Based Indoor Localization: Challenges, Advances, and Comparisons. *ACM Comput. Surv.* 50, 6, Article 97 (December 2017), 37 pages. DOI: <https://doi.org/10.1145/3139222>
- Kirk M. (2017). *Thoughtful Machine Learning with Python: A Test-Driven Approach*, O'Reilly, Chapter 3 K-Nearest Neighbors, pág. 33.
- Larrañaga, P., Inza, I., & Moujahid, A. (1997). Tema 6. Clasificadores Bayesianos. *Departamento de Ciencias de la Computación e Inteligencia Artificial–Universidad del País Vasco-Euskal Herriko Unibertsitatea*.
- Larose, D. T. (2005). k-nearest neighbor algorithm. *Discovering Knowledge in Data: An Introduction to Data Mining*, 90-106.
- Luo C., Hong H., Chan M. C. (2014). Piloc: A self-calibrating participatory indoor localization system. In *Proc. of the 13th International Symposium on Information Processing in Sensor Networks (IPSN-14)*, pp. 143–153.
- Maus, S., Macmillan, S., McLean, S., Hamilton, B., Thomson, A., Nair, M., & Rollins, C. (2010). The US/UK world magnetic model for 2010-2015.

- Mucherino A., Papajorgji P. J., Pardalos P. M. (2009). k-Nearest Neighbor Classification, in *Data Mining in Agriculture*, Springer New York, 2009, pp. 83–106.
- Nerguizian C., Nerguizian V., (2007) Indoor Fingerprinting Geolocation using Wavelet-Based Features Extracted from the Channel Impulse Response in Conjunction with an Artificial Neural Network, *2007 IEEE International Symposium on Industrial Electronics*, Vigo, 2007, pp. 2028-2032. doi: 10.1109/ISIE.2007.4374919
- Pei, L., Zhang, M., Zou, D., Chen, R., & Chen, Y. (2016). A Survey of Crowd Sensing Opportunistic Signals for Indoor Localization. *Mobile Information Systems*, 2016.
- Peng, W., Chen, J., & Zhou, H. (2009). An Implementation of IDE3 Decision Tree Learning Algorithm. *Machine Learning*, 1, 1–20.
- Rai A., Chintalapudi K. K., Padmanabhan V. N., Sen R. (2012). Zee: zero-effort crowdsourcing for indoor localization. In *Proceedings of the 18th annual International Conference on Mobile computing and networking*, pp. 293–304.
- Rokach, L., & Maimon, O. (2005). Decision trees. In *Data mining and knowledge discovery handbook* (pp. 165-192). Springer, Boston, MA.
- Rosado-Gómez A. A., Vergel-Ibáñez A. (2015) Minería de datos aplicada a la demanda del transporte aéreo en Ocaña, Norte de Santander, *Tectura*, 19(25): 101-113.
- Stanton, J. M. (2013). Introduction to data science. Syracuse: Syracuse University .
- Stonehem B., (2016) Google Android Firebase: Learning the Basics, Createspace Independent Publishing Platform, 40 páginas. ISBN: 978-1535004466
- Shu Y., Huang Y., Zhang J., Cuoé P., Cheng P., Chen J., Shin K. G. (2015). Gradient-based Fingerprinting for Indoor Localization and Tracking, *IEEE Trans. Ind. Electron.* 2015(63): 2424–2433.
- Storms W., Shockley J., Raquet J., (2010) Magnetic field navigation in an indoor environment, In *proceedings of 2010 Ubiquitous Positioning Indoor Navigation and Location Based Service*, Kirkkonummi, 2010, pp. 1-10. doi: 10.1109/UPINLBS.2010.5653681

- Taheri A., Singh A., Emmanuel A. (2004). Location fingerprinting on infrastructure 802.11 wireless local area networks (WLANs) using Locus. In *proceedings of 29th Annual IEEE International Conference on Local Computer Networks*, IEEE, pp. 676-683.
- Wirth R., Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pp. 29-39.
- Witten H., Frank E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, pp.
- Wu X., Kumar V., Quinlan J. R., Ghosh J., Yang Q., Motoda H., McLachlan G. J., Ng A., Liu B., Yu P. S., Zhou Z., Steinbach M., Hand D. J., Steinberg D. (2007). Top 10 algorithms in data mining. *Knowledge Information Systems* 14(1): 1-37. DOI=<http://dx.doi.org/10.1007/s10115-007-0114-2>.
- Xie, H., Gu, T., Tao, X., Ye, H., & Lv, J. (2014, September). MaLoc: A practical magnetic fingerprinting approach to indoor localization using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, pp. 243-253.
- Yang S., Dessai P., Verma M., Gerla M. (2013) Freeloc: Calibration-free crowdsourced indoor localization. In *Proceedings IEEE International Conference on Computer Communications (INFOCOM)*, pp. 2481–2489.
- Yanying Gu, Anthony Lo, Ignas Niemegeers. (2009) A survey of indoor positioning systems for wireless personal networks. *IEEE Communications surveys & tutorials*, 11(1):13–32.
- Zhang C., Subbu K. P., Luo J., Wu J. (2015) Groping: Geomagnetism and crowdsensing powered indoor navigation. *IEEE Transactions on Mobile Computing*, 14(2):387–400.
- Zhang X., Jin Y., Tan H. X., Soh W. S. (2014, April). Cimloc: A crowdsourcing indoor digital map construction system for localization. In *Proceedings of 2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, IEEE, pp. 1-6.
- Zhang M., Zhang S., Cao J., Mei, H. (2006) A Novel Indoor Localization Method Based on Received Signal Strength using Discrete Fourier Transform, *2006 First International Conference on Communications and Networking in*

China, Beijing, 2006, pp. 1-5.
doi: 10.1109/CHINACOM.2006.344776

Zheng Z., Chen Y., He. T., Sun L., Chen D. (2015). Feature Learning for Fingerprint-Based Positioning in Indoor Environment, International Journal of Distributed Sensors Networks, pp. 1-11.

Apendice A: Datos de Salida

Código para normalización de los datos R

Normalización de la BD

```
normz=(Nexus_4$norm-mean(Nexus_4$norm))/sd(Nexus_4$norm)
Nexus_4$norm=trunc(rank(normz))/length(normz)
```

```
normz=(Nexus_4$x-mean(Nexus_4$x))/sd(Nexus_4$x)
Nexus_4$x=trunc(rank(normz))/length(normz)
```

```
normz=(Nexus_4$y-mean(Nexus_4$y))/sd(Nexus_4$y)
Nexus_4$y=trunc(rank(normz))/length(normz)
```

```
normz=(Nexus_4$z-mean(Nexus_4$z))/sd(Nexus_4$z)
Nexus_4$z=trunc(rank(normz))/length(normz)
```

Código para algoritmo kNN en R

```
# cargar librerias
library(class)
library(readr)
library(caret)
```

```
#cargar datos
Crowd7 <- read_csv("C:/Users/pablo_000/Dropbox/ISABEL/2017-1/Analisis-
Mapa/CASE CROWD7 MOBILE/Crowd7.csv")
```

```
# definir el 70%/30% training/test del conjunto de datos
```

```
split=0.70
trainIndex <-createDataPartition(Crowd7$place, p=split, list=FALSE)
data_train<-Crowd7[trainIndex,]
data_test<-Crowd7[-trainIndex,]
```

```
# entrenar el modelo de vecinos mas proximos
```

```
#modeloK<-knn(data_train[-5], data_test[-5], data_train$place, k=21,
prob=TRUE)
```

```

modeloK=knn(data_train[-5], data_test[-5], data_train$place, k=9, prob=TRUE)
modeloK7=knn(data_train[-5], data_test[-5], data_train$place, k=7,
prob=TRUE)
modeloK3=knn(data_train[-5], data_test[-5], data_train$place, k=3,
prob=TRUE)

```

```
table(modeloK,data_test$place)
```

```
knn3.matrix()
```

```
x_test=as.factor(data_test$place)
```

```
predicciones<-predict(modeloK7,data_test$place)
```

```

#Matriz de confusion
crowdtab

```

Código para algoritmo J48 en R

```

# cargar librerias
library(RWeka)
library(caret)
library(readr)

#cargar datos
Crowd7 <- read_csv("C:/Users/pablo_000/Dropbox/ISABEL/2017-1/Analisis-
Mapa/CASE CROWD7 MOBILE/Crowd7.csv")

# definir el 70%/30% training/test del conjunto de datos

split=0.70
trainIndex <-createDataPartition(Crowd7$place, p=split, list=FALSE)
data_train<-Crowd7[trainIndex,]
data_test<-Crowd7[-trainIndex,]

# entrenar el modelo de arboles

modeloA<-J48(as.factor(place)~.,data=data_train)
data_test.pred<- predict(modeloA, newdata=data_test)

crowdtab<-table(data_test$place,data_test.pred)

```

```

confusionMatrix(data_test.pred,data_test$place)

# sensibilidad y especificidad

sensitivity(crowdtab,c("LAB A","LAB B", "LAB C","LAB D","LAB E"))
specificity(crowdtab,c("LAB A","LAB B", "LAB C","LAB D","LAB E"))

#Matriz de confusion
crowdtab

```

Código para algoritmo Naïve Bayes en R

```

# cargar librerias
library(RWeka)
library(caret)
library(readr)

#cargar datos
Crowd7 <- read_csv("C:/Users/pablo_000/Dropbox/ISABEL/2017-1/Analisis-
Mapa/CASE CROWD7 MOBILE/Crowd7.csv")

# definir el 70%/30% training/test del conjunto de datos

split=0.70
trainIndex <-createDataPartition(Crowd7$place, p=split, list=FALSE)
data_train<-Crowd7[trainIndex,]
data_test<-Crowd7[-trainIndex,]

# entrenar el modelo de arboles

modeloA<-J48(as.factor(place)~.,data=data_train)
data_test.pred<- predict(modeloA, newdata=data_test)

crowdtab<-table(data_test$place,data_test.pred)

confusionMatrix(data_test.pred,data_test$place)

# sensibilidad y especificidad

```

```
sensitivity(crowdtab,c("LAB A","LAB B", "LAB C","LAB D","LAB E"))  
specificity(crowdtab,c("LAB A","LAB B", "LAB C","LAB D","LAB E"))
```

```
#Matriz de confusion  
Crowdtab
```