

FP-FLOCK: UN ALGORITMO PARA EL DESCUBRIMIENTO DE  
PATRONES DE AGRUPACIÓN DE OBJETOS MÓVILES EN  
BASES DE DATOS ESPACIO TEMPORALES

OMAR ERNESTO CABRERA ROSERO



UNIVERSIDAD DE NARIÑO  
FACULTAD DE INGENIERÍA  
DEPARTAMENTO DE SISTEMAS  
PROGRAMA DE INGENIERÍA DE SISTEMAS  
SAN JUAN DE PASTO  
2014

FP-FLOCK: UN ALGORITMO PARA EL DESCUBRIMIENTO DE  
PATRONES DE AGRUPACIÓN DE OBJETOS MÓVILES EN  
BASES DE DATOS ESPACIO TEMPORALES

OMAR ERNESTO CABRERA ROSERO

TRABAJO DE GRADO PRESENTADO COMO REQUISITO  
PARCIAL PARA OPTAR AL TÍTULO DE INGENIERO DE  
SISTEMAS

DIRECTOR: ANDRES OSWALDO CALDERON ROMERO, MSC.

UNIVERSIDAD DE NARIÑO  
FACULTAD DE INGENIERÍA  
DEPARTAMENTO DE SISTEMAS  
PROGRAMA DE INGENIERÍA DE SISTEMAS  
SAN JUAN DE PASTO  
2014

## **NOTA DE RESPONSABILIDAD**

“Las ideas y conclusiones aportadas en la tesis de grado, son responsabilidad exclusiva de sus autores”.

Artículo 13 del acuerdo N.º 324 del 11 de octubre de 1966, emanado del Honorable Consejo Directivo de la Universidad de Nariño

“La Universidad de Nariño no se hace responsable de las opiniones o resultados obtenidos en el presente trabajo y para su publicación priman las normas sobre el derecho de autor”

Artículo 13, Acuerdo N. 005 de 2010 emanado del Honorable Consejo Académico.

Nota de aceptación:

---

---

---

---

---

Firma del presidente del jurado

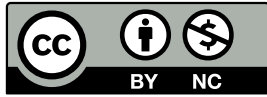
---

Firma del jurado


---


Firma del jurado

San Juan de Pasto, 2014





**You are free:**

 **to Share** – to copy, distribute and transmit this work

 **to Remix** – to adapt this work

**Under the following conditions:**

 **Attribution** – You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work)

 **Noncommercial** – You may not use this work for commercial purposes.

Subject to conditions outlined in the license.

This work is licensed under the *Creative Commons Attribution-NonCommercial 3.0 Unported* License. To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc/3.0/>

or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

*Dedicado a:*

*Mis padres por darme su apoyo incondicional,  
creer en mí y motivarme día a día.*

*Mi hermano y hermana por brindarme su apoyo y compañía.*

*Sandra Cristina Muñoz Castillo (My Elf Princess)  
por haberme entregado su amor, lealtad y  
darme siempre soporte, fuerza y esperanza.*

***“Lucas Poldrosky”***

## **AGRADECIMIENTOS**

Al Sistema de Investigaciones de la Universidad de Nariño, por haber financiado esta investigación.

A Andrés Oswaldo Calderón Romero, mi asesor y amigo, por su disposición, voluntad, compromiso y por haberme despertado el interés en el tema.

A Ricardo Timarán Pereira, por haberme abierto las puertas en el grupo de investigación GRIAS y así poder iniciar mi vida investigativa.

A todos mis amigos y amigas que han estado apoyándome de una u otra manera, en especial a Mateo Guerrero Restrepo, Maria Isabel Gómez y Alicia Guerrero.

A Jairo Guerrero Garcia, por ser mi profesor y mi amigo.

A todos los profesores, por sus enseñanzas y apoyo sincero a lo largo del pregrado.

A toda la comunidad de Software Libre por enseñarme el valor de compartir.

# RESUMEN

El amplio uso de sistemas de localización como dispositivos GPS y RFID junto con el masivo uso dispositivos móviles han hecho que la disponibilidad y acceso a bases de datos espacio temporales se hayan incrementado de manera considerable durante los últimos años. Esta gran cantidad de datos ha motivado el desarrollo de técnicas más eficientes para procesar consultas acerca del comportamiento de los objetos en movimiento, como descubrir patrones de comportamiento entre las trayectorias de objetos móviles durante un periodo continuo de tiempo. Diversos estudios se han centrado en la consulta de patrones que capturan el comportamiento de los diversas entidades en movimiento, los cuales se reflejan en colaboraciones tales como clústers móviles, consulta de convoyes y patrones de agrupamiento. En esta investigación se da a conocer una propuesta para descubrir patrones de agrupamiento, tradicionalmente conocidas como flocks, la cual esta basada en un enfoque de patrones frecuentes. Se presenta el algoritmo FP-Flock para la detección de patrones tanto en línea como fuera de línea. Ambas alternativas fueron comparadas con dos algoritmos del mismo tipo, Basic Flock Evaluation (BFE) y LCMFlock. El desempeño y comportamiento se midió en distintos conjuntos de datos, tanto sintéticos como reales.

**Palabras clave:** patrones de agrupamiento, patrones frecuentes de minería, patrones de movimiento, base de datos espacio-temporal.



# ABSTRACT

The widespread use of location systems such as GPS and RFID along with the massive use of mobile devices have allowed a significantly increasing in the availability and access to spatio-temporal databases in recent years. This large amount of data has motivated the development of more efficient techniques to process queries about the behavior of moving objects, like discovering behavior patterns among trajectories of moving objects over a continuous period of time. Several studies have focused on the query patterns that capture the behavior of the various entities in motion, which are reflected in collaborations such as mobile clusters, convoys query and clustering patterns. In this research, we provided an approach to find grouped patterns, traditionally known as flocks, which is based on an approach of frequent patterns. FP-Flock algorithm is presented for detecting patterns both online and offline are presented. Both alternatives were compared with two algorithms of the same type, Basic Flock Evaluation (BFE) and LCMFlock. The performance and behavior was measured in different datasets, both synthetic and real.

**Keywords:** flock patterns, frequent patterns mining, movement patterns, spatio-temporal databases.

## CONTENIDO

INTRODUCCIÓN	12
REFERENCIAS	15

## ACRÓNIMOS

**BE** Basic Flock Evaluation

**CAD** Computer Aided Design

**DBMS** Database Management System

**FPFlock** Frequent Pattern Flock

**GPS** Global Positioning System

**LCM** Linear time Closed itemset Miner

**LCMFlock** Linear time Closed item set Miner Flock

**MOD** Moving Objects Databases

**RFID** Radio Frequency IDentification

**SUMO** Simulation of Urban MObility

**VLSI** Very Large Scale Integration

## INTRODUCCIÓN

**Planteamiento del problema** En el proceso investigativo realizado en el Grupo de Investigación Aplicada en Sistemas - GRIAS, en la línea de investigación de Herramientas y Sistemas de Gestión de Conocimiento y Recuperación de Información, se han desarrollado dos proyectos de investigación financiados por el sistema de investigaciones de la Universidad de Nariño: uno por la convocatoria estudiantil denominado Construcción de una Ontología de Aplicación que Soporte la Búsqueda Inteligente sobre los Trabajos de Grado de la Universidad de Nariño denominada SAWA, utilizando la herramienta de software libre Protégé"[?] y otro en la convocatoria de trabajos de grado denominada ÜMAYUX: Un Modelo de Software de Gestión de Conocimiento Soportado en una Ontología Dinámica Débilmente Acoplado con un Gestor de Bases de Datos para la Universidad de Nariño"[?]. Estos proyectos fueron delimitados a los trabajos de grado del programa de Ingeniería de Sistemas de la Universidad de Nariño. Como resultado de estos proyectos se cuenta con MASKANA, un prototipo de gestor documental para recuperación de información relacionada con los trabajos de grado del programa de Ingeniería de Sistemas almacenados en formato digital. En estos proyectos se dispone de un repositorio textual de documentos no estructurados sin etiqueta de clase, el cual se limitó a encontrar relaciones semánticas sin tener en cuenta los conceptos y entidades del conocimiento (NER).

"Los conceptos son elementos esenciales para el reconocimiento del mundo que nos rodea. Ellos constituyen una representación de una clase de cosas. Frecuentemente, se suelen confundir o utilizar indistintamente los términos concepto y palabra. El concepto [escuela], por ejemplo, debe ser distinguido de la palabra 'escuela'. [Escuela] es un tipo de [institución educativa]. El concepto [escuela] puede, por ejemplo, ser expresado por las palabras 'escuela', 'lugar para educar', 'institución educativa'. Los conceptos están profundamente relacionados unos con otros de manera que la activación de unos genera la activación de otros. Los vínculos que los interconectan se denominan relaciones conceptuales. Este tipo de relaciones no debe ser confundido con las relaciones entre términos o palabras. Mientras que a las primeras se las suele denominar relaciones conceptuales, a las segundas se las suele denominar relaciones semánticas. Por ejemplo,

las relaciones de sinonimia o de homonimia son relaciones semánticas, mientras que las relaciones taxonómicas y temáticas son relaciones primordialmente entre conceptos"[?].

Se han propuesto diferentes técnicas de minería de textos, en [?] describen los enfoques de extracción de NER y conceptos ligados al conocimiento, en [?] aplican técnicas para extraer palabras clave de documentos textuales. En [?], [?], [?] y [?] proponen aplicar técnicas de minería de textos para representar documentos no estructurados, en [?] y [?] usan grafos conceptuales como representación del contenido de los textos, y obtiene algunos patrones descriptivos de los documentos aplicando varios tipos de operaciones sobre estos grafos.

Estos antecedentes proponen diferentes alternativas de minería de textos pero ninguno de ellos aplicado al dominio de trabajos de grado.

Esto implica investigar diferentes técnicas de minería de textos y minería de datos, aplicarlas en el repositorio, evaluar su correcto funcionamiento e interpretar los patrones obtenidos generando conocimiento útil para el repositorio de la biblioteca Alberto Quijano Guerrero de la universidad de Nariño.

## **Trabajos Relacionados**

Las organizaciones mayormente disponen de información en documentos de texto no estructurado, se encuentran tipificados de esa manera ya que la información contenida en el documento no tiene ningún orden de estructura, ésta información tiene mayor riesgo de no ser encontrada por los buscadores, pues no contiene parámetros establecidos que proporcionen la información que se está buscando y sea presentada al usuario, por esta razón la minería de texto adquiere un rol importante ya que es el proceso de extraer información interesante y conocimiento no trivial de textos no estructurados incluyendo tecnologías para extracción de información, seguimiento de temas, generación automática de resúmenes de textos, categorización, agrupamiento, relaciones entre conceptos, visualización y respuesta automática de preguntas.

[?] describe los principales enfoques de extracción y reconocimiento de NER (entidades con nombre), el NER desempeña un papel muy importante en diversos problemas relacionados a la búsqueda automática y la categorización de textos.

En [?] se propone un nuevo método para la caracterización de documentos que sin importar el idioma en el que el documento esté escrito, permite extraer el conjunto de

palabras clave más adecuado. Su funcionamiento se basa en una Red Neuronal, que luego de ser entrenada es capaz de decidir para cada término del documento si se trata de una palabra clave o no. El ingreso del documento a la Red Neuronal implicó la definición de una representación numérica adecuada que permite medir la participación de un término dentro del documento.

Utilizando las técnicas de minería de texto se pretende obtener una serie de conjuntos de datos estructurados, para poder aplicar algoritmos de aprendizaje automático como se lo propone en [?], [?], [?] y [?] logrando una categorización y clasificación de documentos adecuada, [?] propone un método que relaciona e integra técnicas de procesamiento de lenguaje natural, agrupamiento (clustering) y modelos de Markov como una solución de bajo costo, dependiente del dominio, para la evaluación automática de la organización en textos argumentativos.

Otra propuesta de utilización de minería de texto es la de Grobelnik, Mladenic and Jermol [?], en la cual se pretende potenciar una aplicación de construcción de ontologías/taxonomía a partir de un conjunto de documentos planos, realizar búsquedas en la base de documentos y tratar problemas específicos del lenguaje, por su parte [?],[?] proponen sistemas para resumir textos, agrupar documentos e interpretar el conocimiento de los grupos obtenidos para una fácil compresión por parte del usuario.

Arco et al.[?] estudia el impacto de la representación del texto en el ámbito de la clasificación no supervisada (CNS) de documentos. Tomando como referencia una representación basada en un modelo de espacio vectorial de términos, se analizan diferentes técnicas de representación de los datos sobre espacios de menor dimensionalidad (obtenidas mediante técnicas de extracción de términos como el Análisis de Semántica Latente, la Factorización en Matrices No Negativas y el Análisis en Componentes Independientes) para mejorar la CNS de un corpus de documentos.

En [?] y [?] emplean minería de texto para la semejanza entre estructuras semánticas usando grafos conceptuales como representación del contenido de los textos, y obtiene algunos patrones descriptivos de los documentos aplicando varios tipos de operaciones sobre estos grafos.

## **Objetivos**

### **Objetivo General**

Descubrir relaciones conceptuales entre los trabajos de grado de la Universidad de Nariño (Colombia) utilizando técnicas de minería de texto que facilite la recuperación de trabajos de grado relacionados con la temática de la búsqueda identificando similitudes y diferencias entre ellos.

### **Objetivos específicos**

- Apropiar el conocimiento en algoritmos de minería de texto, algoritmos de agrupación y aprendizaje automático.
- Construir, limpiar y transformar el repositorio de documentos de trabajos de grado de la universidad de Nariño.
- Implementar los algoritmos de minería de texto seleccionados en la herramienta MASKANA.
- Descubrir las relaciones conceptuales entre los trabajos de grado y evaluar los resultados.
- Elaboración del documento final de tesis.

### **Organización de la tesis**

En el capítulo 2 se elabora un marco teórico referente a minería de textos. En el capítulo 3 se presenta la construcción, limpieza, transformación del corpus de documentos de trabajos de grado de la universidad de Nariño y los experimentos realizados. En el capítulo 4 se presentan los resultados. En el capítulo 5 se muestra las conclusiones y trabajos futuros en base a los resultados.