INSTITUTO POLITÉCNICO NACIONAL

Centro de Investigación en Computación

TESIS

Análisis y clasificación de la publicación científica del CIC

QUE PARA OBTENER EL GRADO DE:

MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

Ing. Oscar Alberto Rocha Arcos

DIRECTORES DE TESIS:

Dr. Gilberto Lorenzo Martínez Luna

Dr. Adolfo Guzmán Arenas

México, Ciudad de México a 12 de mayo de 2022

SIP-13
REP 2017

INSTITUTO POLITÉCNICO NACIONAL
SECRETARIA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REGISTRO DE TEMA DE TESIS
Y DESIGNACIÓN DE DIRECTOR DE TESIS

Ciudad de México, a 12 de mayo del 2022

El Colegio de Profesores de Posgrado del **Centro de Investigación en Computación** en su Sesión
(Unidad Académica)

Ordinaria No 9 celebrada el día 30 del mes septiembre de 2021, conoció la solicitud presentada por el (la) alumno (a):

Apellido Paterno:	ROCHA	Apellido Materno:	ARCOS	Nombre (s):	OSCAR ALBERTO

Número de registro: A 2 0 0 4 2 7

del Programa Académico de Posgrado: **Maestría en Ciencias de la Computación**

Referente al registro de su tema de tesis; acordando lo siguiente:

1.- Se designa al aspirante el tema de tesis titulado:

"Análisis y clasificación de la publicación científica del CIC"

Objetivo general del trabajo de tesis:

Se trata de conocer la relevancia de los trabajos del CIC, expresados en las tesis de sus alumnos, en el entorno tecnológico, innovación e investigación en computación, y en las áreas que incide. Proyecto estratégico nacional (PRONACE) al que pertenece: Educación.

2.- Se designa como Directores de Tesis a los profesores:

Director: **Dr. Gilberto Lorenzo Martínez Luna** 2° Director: **Dr. Adolfo Guzmán Arenas**

No aplica: ☐

3.- El Trabajo de investigación base para el desarrollo de la tesis será elaborado por el alumno en:

Centro de Investigación en Computación

que cuenta con los recursos e infraestructura necesarios.

4.- El interesado deberá asistir a los seminarios desarrollados en el área de adscripción del trabajo desde la fecha en que se suscribe la presente, hasta la aprobación de la versión completa de la tesis por parte de la Comisión Revisora correspondiente.

Director(a) de Tesis

Dr. Gilberto Lorenzo Martínez Luna

2º Director de Tesis

Dr. Adolfo Guzmán Arenas

Aspirante

C. Oscar Alberto Rocha Arcos

Presidente del Colegio

Dr. Francisco Hiram Calvo Castro

INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN
DIRECCIÓN

SIP-14
REP 2017

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México siendo las 17:00 horas del día 12 del mes de mayo del 2022 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Posgrado de: Centro de Investigación en Computación para examinar la tesis titulada:

"Análisis y clasificación de la publicación científica del CIC"

del (la) alumno (a):

Apellido Paterno:	ROCHA	Apellido Materno:	ARCOS	Nombre (s):	OSCAR ALBERTO

Número de registro: A 2 0 0 4 2 7

Aspirante del Programa Académico de Posgrado: **Maestría en Ciencias de la Computación**

Una vez que se realizó un análisis de similitud de texto, utilizando el software antiplagio, se encontró que el trabajo de tesis tiene 06 % de similitud. **Se adjunta reporte de software utilizado**.

Después que esta Comisión revisó exhaustivamente el contenido, estructura, intención y ubicación de los textos de la tesis identificados como coincidentes con otros documentos, concluyó que en el presente trabajo **SI** ☐ **NO** ☒ **SE CONSTITUYE UN POSIBLE PLAGIO.**

JUSTIFICACIÓN DE LA CONCLUSIÓN: *(Por ejemplo, el % de similitud se localiza en metodologías adecuadamente referidas a fuente original)*
Se utilizan términos comunes encontrados en las fuentes

****Es responsabilidad del alumno como autor de la tesis la verificación antiplagio, y del Director o Directores de tesis el análisis del % de similitud para establecer el riesgo o la existencia de un posible plagio.**

Finalmente y posterior a la lectura, revisión individual, así como el análisis e intercambio de opiniones, los miembros de la Comisión manifestaron **APROBAR** ☒ **SUSPENDER** ☐ **NO APROBAR** ☐ la tesis por **UNANIMIDAD** ☒ o **MAYORÍA** ☐ en virtud de los motivos siguientes:
Cumple con los requisitos de una tesis de Maestría en Ciencias de la Computación

COMISIÓN REVISORA DE TESIS

Dr. Gilberto Lorenzo Martínez Luna
Director de Tesis

Dr. Grigori Sidorov

Dr. René Luna Garcia

Dr. Adolfo Guzmán Arenas
2° Director de Tesis

Dr. Jesús Manuel Olivares Ceja

Dr. Salvador Godoy Calderón

INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN
DIRECCIÓN
IPN-CIC

Dr. Francisco Hiram Calvo Castro
PRESIDENTE DEL COLEGIO DE PROFESORES

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA DE AUTORIZACIÓN DE USO DE OBRA PARA DIFUSIÓN

En la Ciudad de México el día **18** del mes de **Junio** del año **2022**, el que suscribe **Oscar Alberto Rocha Arcos** alumno del programa **Maestría en Ciencias de la Computación** con número de registro **A200427**, adscrito a **Centro de Investigación en Computación** manifiesta que es autor(a) intelectual del presente trabajo de tesis bajo la dirección de **Dr. Gilberto Lorenzo Martínez Luna** y **Dr. Adolfo Guzmán Arenas** y cede los derechos del trabajo intitulado **"Análisis y clasificación de la publicación científica del CIC"** al Instituto Politécnico Nacional, para su difusión con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expresado del autor y/o director(es). Este puede ser obtenido escribiendo a las siguiente(s) dirección(es) de correo. **alberto.oscar96@gmail.com** Si el permiso se otorga, al usuario deberá dar agradecimiento correspondiente y citar la fuente de este.

Oscar Alberto Rocha Arcos

Nombre completo y firma autografá del (de la) estudiante

AGRADECIMIENTOS

Agradecer a mis padres, por estar siempre apoyándome incondicionalmente. Por todo el amor y paciencia que me han brindado, así como sus consejos y motivación que sirvieron para la conclusión de este trabajo.

A mis directores de tesis, el Dr. Gilberto Lorenzo Martínez Luna y al Dr. Adolfo Guzmán Arenas, por su orientación, confianza, apoyo, críticas y comentarios que me hicieron durante la realización de este trabajo, siendo gran inspiración durante este periodo y el futuro venidero.

Agradecer a mi comité tutorial por compartir sus conocimientos y experiencia ayudando a mejorar este trabajo y mi persona.

Agradecer a mis compañeros del CIC, que me apoyaron con sus sugerencias, comentarios y consejos, y me hicieron más amena mi estancia.

Agradecer al Centro de Investigación en Computación, y todos lo que lo conforman por brindarnos herramientas, soporte y atención durante mi estancia en el CIC.

Agradezco a las instituciones que me apoyaron a realizar mis estudios de maestría, el Instituto Politécnico Nacional y al Consejo Nacional de Ciencia y Tecnología.

Oscar Alberto Rocha Arcos.

Acrónimos

CIC	Centro de Investigación en Computación
IPN	Instituto Politécnico Nacional
ACM	Association for Computing Machinery (Asociación de Maquinaria Computacional)
I+D	Investigación y Desarrollo
ISI	Institute for Scientific Information (Instituto de Información Científica)
SCI	Science Citation Index (Índice de Citas de la Ciencia)
IEEE	Institute of Electrical and Electronics Engineers (Instituto de Ingenieros Eléctricos y Electrónicos)
NLP	Natural Langauge Processing (Procesamiento de Lenguaje Natural)
IA	Inteligencia Artificial
TF	Term Frequency (Frecuencia de término)
DF	Document Frequency (Frecuencia de documento)
IDF	Inverse Document Frequency (Frecuencia inversa de documento)
TF-IDF	Term Frequency – Inverse Document Frequency
BoW	Bag of Words (Bolsa de Palabras)
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
UML	Unified Modeling Language (Lenguaje de Modelado Unificado)
PDF	Portable Document File (Formato de archivo "Formato de Documento Portátil")
TXT	Formato de archivo que contiene texto plano

DAT	Formato de archivo que puede contener cualquier información
CSV	Comma Separated Values (Formato de archivo "Valores Separados por Comas")
LPC	Linear Predictive Coding (Codificación predictiva lineal)
PDS	Procesador Digital de Señales
DSP	Digital Signal Processor (Procesador digital de señales)
MEMS	Microelectromechanical systems (Sistemas microelectromecánicos)
HCI	Human Centered Interface (Interfaz centrada al humano)
TP	True Positive (Verdaderos Positivos)
TN	True Negative (Verdaderos Negativos)
FP	False Positive (Falsos Positivos)
FN	False Negative (Falsos Negativos)
F1	Una medida que combina precisión y recuperación es la media armónica de precisión y recuperación.

ÍNDICE DE CONTENIDO

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

1. Definición del proyecto

Acrónimos del capítulo

CIC	Centro de Investigación en Computación
IPN	Instituto Politécnico Nacional
ACM	Association for Computing Machinery (Asociación de Maquinaria Computacional)
I+D	Investigación y Desarrollo

RESUMEN

Actualmente se generan elevados volúmenes de información derivados de la actividad científica, aunado a esto, la evolución del conocimiento científico en todas sus áreas de investigación y específicamente en las áreas de la computación y afines, está en constante transformación. Partiendo de esto, se tiene la necesidad de analizar y esquematizar esta información generada, para obtener conclusiones sobre su avance, evolución y generación.

El presente documento realiza un análisis y evaluación, de la(s) temática(s) dominantes(s) en las publicaciones del CIC-IPN que se han desarrollado hasta el 2020. Las 690 tesis fueron obtenidas de los repositorios digitales del IPN. Para la evaluación del sistema, 100 tesis se clasificaron de manera manual previamente con base a la ontología multi-jerárquica que sugiere ACM (Association for Computing Machinery, por sus siglas en inglés) para el campo de la computación, sirviendo como Golden Standard. La clasificación de las tesis es utilizando el sistema desarrollado CLASSONTO, que permite clasificar basándose en un árbol de conocimiento. El sistema ocupa un votación y propagación del voto, para la identificación de temáticas, además de un procesamiento de textos. El árbol de conocimiento en el que se basa es enriquecido, por lo que lo convierte en un sistema de mejora continua.

Los resultados obtenidos muestran las clasificaciones de cada tesis, permitiendo inferir que temas trata cada una. De igual manera se hace un análisis y evaluación de los resultados con respecto al tiempo, visualizando el desarrollo de temáticas a través del tiempo.

Esta investigación es de particular interés, pues servirá como una referencia para evaluar la producción científica del CIC, fomentando su estudio y catalogación.

Palabras clave: *clasificación jerárquica, ontologías, modelado de temas.*

Clasificación ACM: *8.5.5.8. Clustering and classification, 11.3.1. Natural Language Processing, 12.8.2. Publishing, 11.4.1.2.4. Topic modelling, 8.5.1.6. Ontologies*

ABSTRACT

Currently high volumes of information derived from scientific activity are generated, coupled with this, the evolution of scientific knowledge in all its research areas and specifically in the areas of computing and related, is in constant transformation. Based on this, it is necessary to analyze and outline this generated information, to obtain conclusions about its progress, evolution, and generation.

This document carries out an analysis and evaluation of the dominant thematic(s) in the CIC-IPN publications that have been developed until 2020. The 690 theses were obtained from the IPN's digital repositories. For the evaluation of the system, 100 theses were previously manually classified based on the multi-hierarchical ontology suggested by ACM (Association for Computing Machinery) for the field of computing, serving as Golden Standard. The classification of the theses is using the system developed CLASSONTO, which allows to classify based on a tree of knowledge. The system occupies a vote and propagation of the vote, for the identification of topics, in addition to a text processing. The knowledge tree on which it is based is enriched, so it becomes a system of continuous improvement.

The results obtained show the classifications of each thesis, allowing to infer which topics each one deals with. In the same way, an analysis and evaluation of the results with respect to time is made, visualizing the development of themes over time.

This research is of particular interest, as it will serve as a reference to evaluate the scientific production of the CIC, promoting its study and cataloguing.

Keywords: *hierarchical classification, ontologies, topic modelling.*

ACM Classification: *8.5.5.8. Clustering and classification, 11.3.1. Natural Language Processing, 12.8.2. Publishing, 11.4.1.2.4. Topic modelling, 8.5.1.6. Ontologies*

1.1. INTRODUCCIÓN

A nivel mundial, la necesidad de aplicar sistemas de evaluación en materia de investigación se ha extendido entre las administraciones públicas. A partir de los resultados, cada gobierno puede establecer mecanismos para optimizar los recursos disponibles y reorientar políticas e instrumentos que enmarcan el desarrollo de la actividad investigadora [1]. Diferentes autores establecen dos tipos de indicadores para la evaluación de sistemas de ciencia y tecnología: los indicadores de insumo y los indicadores de producción científica. Los primeros caracterizan el comportamiento de países o regiones en materia de financiación y recursos humanos, como información de referencia para el desarrollo de la investigación. El segundo grupo está orientado a la evaluación de resultados por medio de publicaciones científicas, como muestra de la consolidación del sistema y su contribución al desarrollo económico y social de un país [2].

Frente al primer grupo de indicadores, a partir de la década de los años sesenta del siglo pasado, la *Organización para la Cooperación y el Desarrollo Económico* (OCDE)[1] y la *Organización de Naciones Unidas para la Educación, la Ciencia y la Cultura* (Unesco)[2] han desarrollado diferentes metodologías, aceptadas por la comunidad internacional, para evaluar la inversión en investigación y desarrollo (I+D) y los recursos humanos que participan en el desarrollo científico. Dichas evaluaciones son de libre acceso a través de los portales estadísticos de las organizaciones mencionadas.

Para analizar las publicaciones científicas, se han desarrollados ciencias (bibliometría, cienciometría, infometría, etc.), para evaluar el desarrollo científico. Ya sea utilizando elementos comunes presentes en todas las publicaciones científicas: referencias bibliográficas, autores, etc., que sirven como indicadores. Sin embargo, desde la creación de la cienciometría por la obra de *Derek John de Solla Price "The advent of science indicators"* el proceso científico ha ido en incremento; desde el número de artículos publicados hasta centros de investigación, revistas, gastos, científicos, patentes, derechos de autor, disertaciones o tesis. Así también la creación de nuevos indicadores cienciométricos o elementos que brinden más contexto para comprender mejor el proceso científico y que disten de los usuales indicadores bibliométricos. Ejemplos más recientes de indicadores cienciométricos son los que se han enfocado en el crecimiento de patrones o tendencias en determinadas disciplinas:

[1] http://www.oecd.org/publications/frascati-manual-2015-9789264239012-en.htm

[2] http://uis.unesco.org/sites/default/files/documents/fs42-global-investments-in-rd-2017-en.pdf

Un ejemplo de un nuevo enfoque es como lo hizo *Trends in the investigation of social determinants of health: selected themes and methods* [3] analizando tendencias en el desarrollo de temas en el ámbito de la salud, obtenidas de la base de datos PubMed que cubre el período 1985-2007, usando como indicador cienciométrico las temáticas de las publicaciones.

Así es como nuevos indicadores cienciométricos han ayudado en la esquematización del conocimiento, para evaluar el desarrollo de la actividad científica.

1.2. Planteamiento del problema

El análisis de las publicaciones científicas constituye un eslabón fundamental dentro del proceso de investigación [4], y para dicho análisis existen ciencias como la bibliometría y la cienciometría. La cienciometría permite el uso de nuevos indicadores cienciométricos para la evaluación de la actividad científica [5], [6], pudiendo ser estos el análisis de las temáticas desarrolladas en las publicaciones científicas. Una revisión de un estudio usando la cienciometría aplicada, entrega como resultado las aristas en el cambio específico de la ciencia, a partir de los temas desarrollados [7].

En México, en muchos centros de investigación no se tienen elementos organizados de tal manera que permitan analizar su producción científica que sea más allá de realizar estadísticas sobre esta [8]. Un análisis más profundo, sería el poder describir las temáticas que desarrolla la producción científica a través del tiempo. Aunado a esto, también está la falta de uso de estándares, criterios o sistemas de clasificación oficiales que permitan categorizar de manera detallada la publicación científica en áreas especializadas, especialmente en las que están en constante transformación y crecimiento como lo es la Computación.

Por ello, esta investigación pretende analizar y evaluar la producción científica del CIC a través de las temáticas(s) dominante(s), basadas en el sistema de clasificación propuesto por ACM para el campo de la computación.

1.3. Objetivos

1.3.1. *Objetivo general*

Diseñar un algoritmo que permita clasificar las publicaciones del CIC-IPN con una clasificación oficial (sistema de clasificación computacional del ACM) para analizarlas y evaluar el desarrollo de las temáticas a través del tiempo.

1.3.2. *OBJETIVOS PARTICULARES*

- Obtención de un árbol de conocimiento a partir de la clasificación computacional establecida por ACM.
- Desarrollo de un algoritmo capaz de clasificar la publicación científica del CIC con base a un árbol de conocimiento.
- Enriquecimiento del árbol de conocimiento a partir de la publicación analizada.
- Visualizar las temáticas que ha tenido la publicación científica del CIC a través del tiempo.

1.4. JUSTIFICACIÓN

El interés de esta investigación está en clasificar las publicaciones del CIC (tesis) por medio de las temáticas desarrolladas, bajo un sistema oficial para su posterior análisis. Esto es de gran importancia debido a que uno de los objetivos del Centro de Investigación en Computación CIC – IPN es divulgar los resultados de su producción científica por medio de documentación escrita.

Es en este punto, donde hay que destacar la contribución particular del análisis por medio de las temáticas, usándolas como indicador cienciométrico, para evaluar la producción científica, y usando un esquema de clasificación oficial, permitirá su fácil comparación. También está la contribución del área de la computación, específicamente: la ciencia de datos y el procesamiento del lenguaje natural, que permiten la adquisición, selección y preparación de corpus de texto para un análisis automático o semiautomático.

Por lo anterior, en un centro de investigación puede ser de gran utilidad hacer una revisión de su producción científica, para proveer de una perspectiva al CIC para la visualización de los temas que fueron y son tendencia, en el ámbito de la computación que siempre se transforma.

1.5. PROPUESTA DE SOLUCIÓN

Se propone una investigación que clasifique la producción científica del CIC (tesis) bajo la clasificación oficial propuesta por el ACM, para la evaluación del desarrollo científico del instituto a través del tiempo.

La Ilustración 1-1 muestra de manera general un diagrama a bloques del sistema final, a continuación, se describe cada etapa del diagrama.

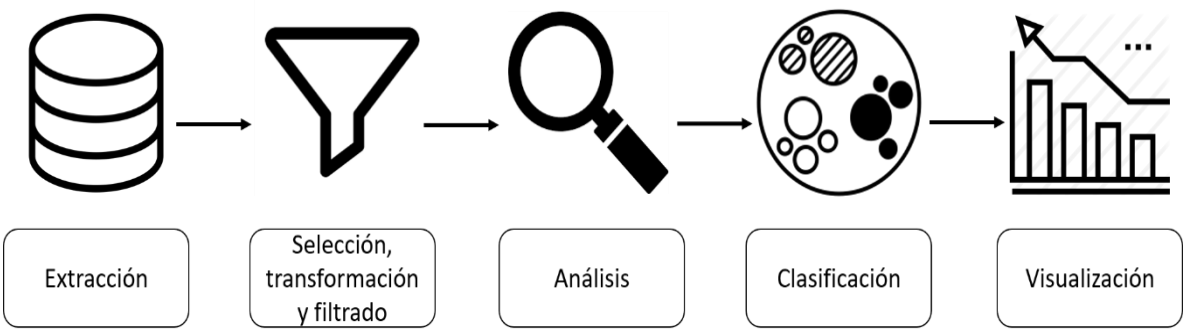


Ilustración 1-1. Diagrama a bloques del prototipo final.

1.5.1. *EXTRACCIÓN*

Se hará una búsqueda y delimitación de fuentes de información en donde se obtengan las tesis del CIC de manera digitalizada.

1.5.2. *SELECCIÓN, TRANSFORMACIÓN Y FILTRADO*

Todas las publicaciones deben de haber sido desarrolladas en el CIC, es decir, el autor debió haber sido estudiante de maestría o doctorado del CIC.

Para la investigación, la información obtenida deberá contener las siguientes características:

- Autor (también deberá contener a los directores de tesis)
- Título
- Resumen
- Fecha
- Documento completo

1.5.3. *ANÁLISIS Y CLASIFICACIÓN*

El análisis se hará usando las publicaciones, para compararlas con el árbol de conocimiento basado en la ontología multi-jerárquica de ACM (Ilustración 1-2), para su posterior clasificación.

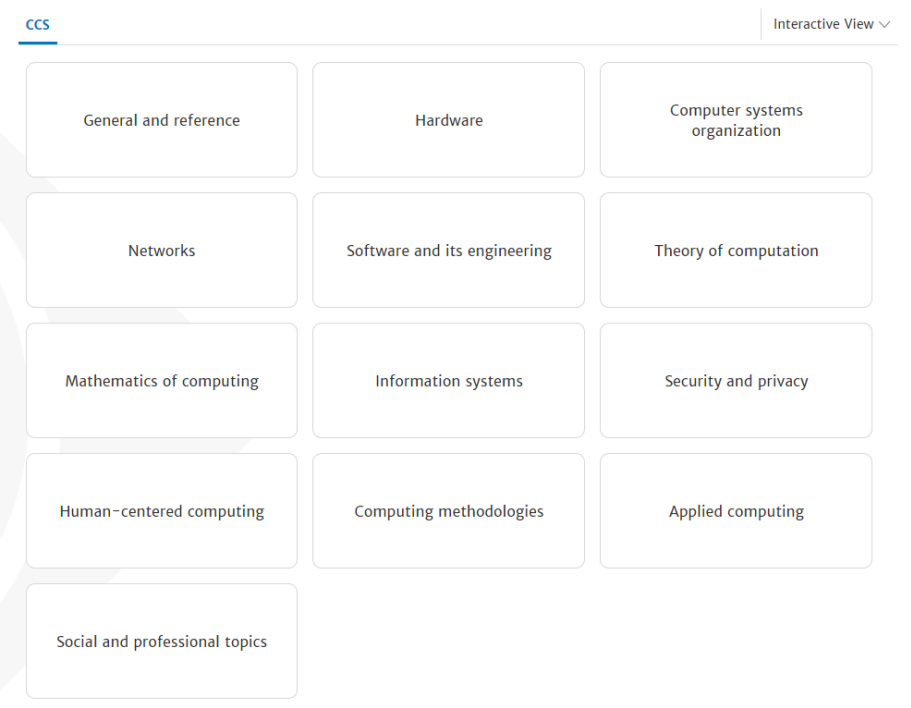


Ilustración 1-2. Las 13 categorías propuestas por ACM para el campo de la computación [3].

1.5.4. *VISUALIZACIÓN Y EVALUACIÓN*

- Visualización:

Se visualizará las categorías (basadas en ACM) con las que se clasificó las publicaciones científicas ponderándolas por la importancia que tienen en el documento: alta, media o baja.

- Evaluación:

Se evaluará el desarrollo de las temáticas (categorías de ACM) en las publicaciones a lo largo del tiempo; pretendiendo responder los siguientes cuestionamientos:

  - ¿Cuáles son las clasificaciones del ACM desarrolladas en el CIC que se les da más continuidad (se vuelven a usar)?
  - ¿Qué clasificaciones del ACM desarrolladas en el CIC tienen una fuerte producción en los últimos años?

[3] https://dl.acm.org/ccs

- ¿Qué clasificaciones del ACM desarrolladas en el CIC están en el olvido?
- ¿Cuáles son las clasificaciones del ACM desarrolladas en el CIC son nuevas?
- ¿Cuáles son las clasificaciones del ACM que más se han desarrollado por año en el CIC?
- ¿Cómo ha sido el desarrollo a través del tiempo de las clasificaciones del ACM históricamente más usadas en el CIC?

1.6. Beneficios esperados y Aportaciones

- Generación de un sistema que sea capaz de clasificar las temáticas en el ámbito de la computación bajo esquemas oficiales (árboles de conocimiento).
- Visualizar el desarrollo a través del tiempo de las temáticas desarrolladas
- Incentivar la clasificación de la publicación científica, bajo esquemas oficiales.
- Facilitar la comprensión de las temáticas que desarrolla el instituto.

1.7. Alcances y Limitaciones

- La publicación científica analizada será con las tesis de maestría y doctorado que se desarrollaron en el CIC.
- El árbol de conocimiento se podrá enriquecer, permitiendo que el sistema sea de mejora continua.
- El sistema no contempla el análisis de la(s) temática(s) en publicaciones con idiomas diferentes al español y al inglés.
- El sistema trabajará con obras completas, por lo que si solo se cuentan con los metadatos o partes de la información (por ejemplo, registros de tesis con solo título y resumen) no se contemplan.
- El periodo de análisis de las publicaciones se comprende hasta el 2020.

1.8. Organización del documento

El presente documento está dividido en seis capítulos a través de los cuales se desarrollarán los diferentes temas relacionados con el trabajo de tesis.

El capítulo dos, Estado del Arte, menciona de manera breve los trabajos previos similares a la investigación.

El capítulo tres, Marco Teórico, toca el tema principal de esta investigación e introduce al lector en los temas básicos de la cienciometría, procesamiento de lenguaje natural y representación del conocimiento.

El capítulo cuatro, Análisis y Diseño, se presenta la metodología utilizada incluyendo la descripción de las fuentes de información, las herramientas, los niveles y unidades de análisis y observación, las variables de estudio, así como el preprocesamiento de los datos y los indicadores a partir de los cuales se realiza esta investigación.

El capítulo cinco, Desarrollo y Resultados, expone los resultados obtenidos y busca dar respuestas a las preguntas planteadas.

Finalmente, el capítulo seis, Conclusiones y Trabajo a Futuro, comprende la discusión, conclusiones obtenidas finalizado el trabajo, y las futuras líneas de investigación que se derivan de este trabajo.

1.9. REFERENCIAS DEL CAPÍTULO

[1] H. F. Moed, "New developments in the use of citation analysis in research evaluation," *Arch. Immunol. Ther. Exp. (Warsz).*, vol. 57, no. 1, pp. 13–18, 2009, doi: 10.1007/s00005-009-0001-5.

[2] J. P. Man, J. G. Weinkauf, M. Tsang, and D. D. Sin, "Why do some countries publish more than others? An international comparison of research funding, English proficiency and publication output in highly ranked general medical journals," *Eur. J. Epidemiol.*, vol. 19, no. 8, pp. 811–817, 2004, doi: 10.1023/B:EJEP.0000036571.00320.b8.

[3] R. K. Celeste, J. L. Bastos, and E. Faerstein, "Trends in the investigation of social determinants of health: selected themes and methods," *Cad. Saude Publica*, 2011, doi: 10.1590/s0102-311x2011000100019.

[4] C. Rueda-Clausen, C. Villa-Roel, and C. Rueda-Clausen, "Indicadores bibliométricos: origen, aplicación, contradicción y nuevas propuestas," *MedUNAB*, vol. 8, no. 1, pp. 29–36, 2005.

[5] J. A. Araújo Ruiz and R. Arencibia Jorge, "Infometría, bibliometría y cienciometría: aspectos teórico-prácticos," *ACIMED*, vol. 10, no. 4, pp. 5–6, 2002, Accessed: Mar. 20, 2021. [Online]. Available: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352002000400004&lng=es&nrm=iso&tlng=es.

[6] A. J. Lotka, "The frequency distribution of scientific productivity," *J. Washingt. Acad. Sci.*, vol. 16, no. 12, pp. 317–323, 1926.

[7] F. A. Poveda Aguja, E. O. Cruz, G. Barajas, C. J. Cabanzo, and COLCIENCIAS, *Introduction to Scientometrics, innovation and scientific activity*, no. June 2019. 2017.

[8] M. Vazquez Gallo, "Análisis cienciométrico en el desarrollo de temas de tesis de maestría en el periodo 1996-2007. Caso de estudio: 'Centro de Investigación en Computación del Instituto Politécnico Nacional,'" 2010.

2. Estado del arte

Acrónimos del capítulo

ISI	Institute for Scientific Information (Instituto de Información Científica)
SCI	Science Citation Index (Índice de Citas de la Ciencia)
IEEE	Institute of Electrical and Electronics Engineers (Instituto de Ingenieros Eléctricos y Electrónicos)
CIC	Centro de Investigación en Computación
IPN	Instituto Politécnico Nacional
ACM	Association for Computing Machinery (Asociación de Maquinaria Computacional)

2.1. INTRODUCCIÓN

En este capítulo se describen de manera breve los trabajos relacionados a la esquematización del conocimiento para la evaluación de diferentes ámbitos específicos y la aplicación de la cienciometría para conocer el volumen y la visibilidad de los resultados de la actividad investigadora [1]. En los siguientes trabajos mencionados, se explican de manera breve las diversas técnicas y herramientas de software que se usaron para la realización de este tipo de estudios cienciométricos.

2.1.1. *MAPPING THE WORLD OF BIOMEDICAL ENGINEERING*

Estudio publicado en 1986 por Eugene Garfield [2], considerado el fundador de la bibliometría y la cienciometría. De igual manera fundó el "Instituto de Información Científica (ISI, por sus siglas en inglés) y el "Índice de Citas de la Ciencia (SCI, por sus siglas en inglés), que permite calcular los factores de impacto, siendo una medida en la importancia de las revistas científicas.

En este estudio se emplea la frecuencia de citas y co-citas para la construcción de agrupamientos (clústeres) que representan los diversos temas de investigación. Este proceso se realiza determinando el número de veces que un artículo es citado por otros artículos a lo largo del tiempo. Dado artículos altamente citados, es importante conocer que tan frecuentemente pares de ellos so co-citados en otros artículos. Teniendo esta información, se generan agrupaciones iniciales, a partir de las cuales se identifican los temas de investigación con mayor número de publicaciones.

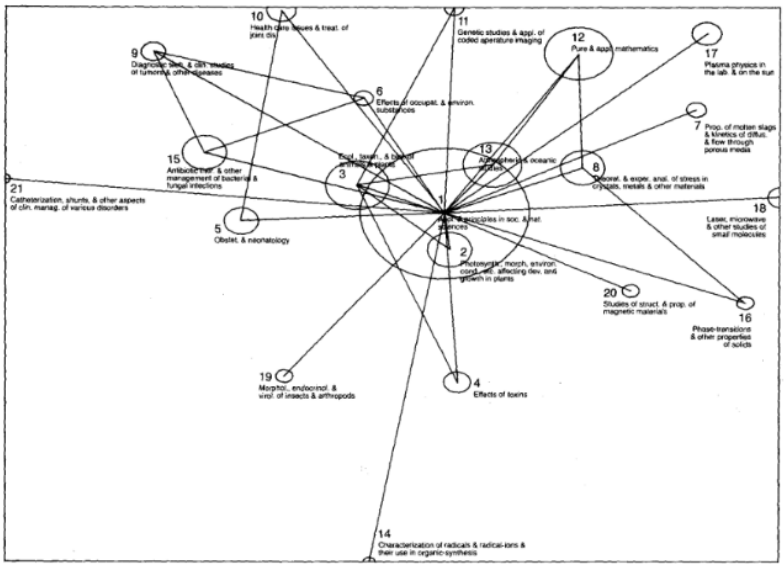


Ilustración 2-1. Mapa que muestra las áreas de investigación mayores en 1984 [2].

2.1.2. *VISUALIZING THE SCIENTIFIC WORLD AND ITS EVOLUTION*

Estudio publicado en 2006 por I. Samoylenko, et al. [3], en el que proponen un enfoque para visualizar el mundo científico y su evolución mediante la construcción de árboles de expansión mínimos y modelando mapas de revistas científicas de dos dimensiones. Para ello evalúan la similitud o "similaridad" entre revistas, formando una matriz de similitud Para visualizar esta similitud, se plasman en un mapa de dos dimensiones en donde las revistas cercanas entre ellas son más similares, caso contrario con revistas alejadas entre sí. Por último, se construye el árbol de expansión mínima, empleando el algoritmo de Kruskal conectando objetos vecinos más cercanos a partir de un grafo inicialmente desconectado. Este árbol Ilustración 2-2, puede representar agrupaciones de las temáticas que tratan las revistas, y sus relaciones entre ellas.

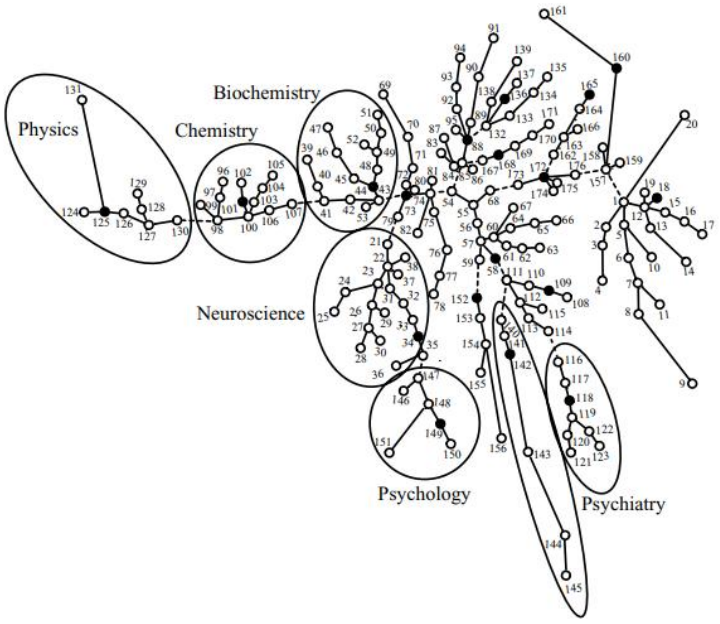


Ilustración 2-2. Árbol que representa agrupaciones del conocimiento científico tratado por revistas [3].

2.1.3. *MODELOS DE DESARROLLO DEL HARDWARE Y SOFTWARE BASADOS EN EL ESTUDIO DE COMPUTACIÓN PARALELA*

Estudio publicado en 2007 por Alejandro Ruiz y Pavel Makagonov [4] el cual permite cuantificar y evaluar la actividad científica en las áreas de electrónica y computación. Para esto proponen un método de construcción de modelos cuantitativos específicos y generales. Siendo el modelo específico, una curva S reconstruida a partir de los 23000 resúmenes recopilados y el modelo general siendo la curva tangente a este conjunto de curvas. Al realizar el estudio, se consideró que cada sistema posee un

ciclo de vida, el cual se ajusta a la curva S, donde se comienza a partir de un desarrollo lento, que posteriormente se incrementa con rapidez hasta llegar a un punto en donde dicho incremento se reduce drásticamente, y si en dado caso aparece un decrecimiento, se puede asociar a la aparición de nuevos sistemas. Este comportamiento se muestra en la Ilustración 2-3. Donde hay una tendencia creciente hasta 1997, y en 1998 disminuye el interés por los temas de computación, habiendo una recuperación del interés en el año 2000 que aumenta gradualmente hasta el 2004.

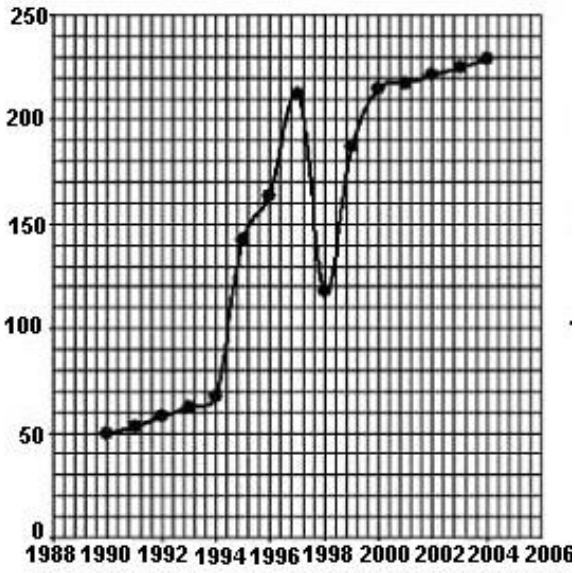


Ilustración 2-3. Histograma de los 2300 artículos de la librería digital de IEEE para los temas de computación paralela, concurrente, distribuida y simultánea, para el período de 1990 al 2004 [4].

2.1.4. *DESARROLLO DE UN SISTEMA DE ANÁLISIS TEMÁTICO DEL CONOCIMIENTO CIENTÍFICO*

Tesis publicada en 2009 por Eduardo Godínez [5] en el CIC-IPN, en donde desarrolla una herramienta de software que permite identificar tendencias, que describen la evolución en una disciplina del conocimiento científico cuyos recursos de información se encuentran ya clasificados por la librería digital de ACM, en sus primeros tres niveles. Las tendencias en las categorías o clasificaciones se identifican mediante la formulación de preguntas dirigidas conocidas como modelos y dependiendo del modelo utilizado, los resultados son representados en gráficas que muestran las tendencias o mapas de conocimiento. Con estas gráficas se puede mostrar la contribución en la producción de los temas o categorías de la disciplina en un lapso, y por consecuencia su popularidad, ya sea si decrecen o incrementan, como se muestra en la Ilustración 2-4.

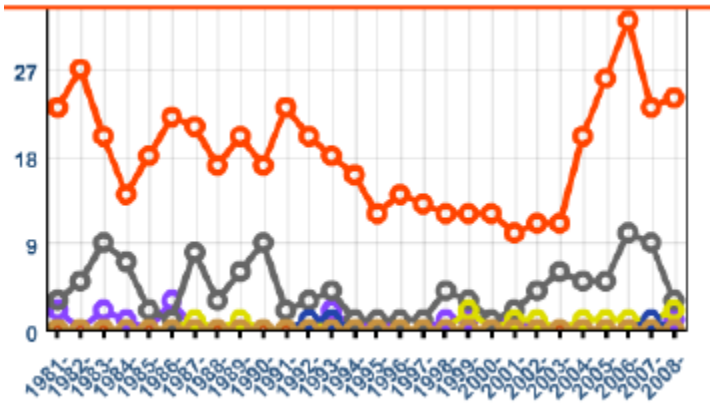


Ilustración 2-4. Publicación de artículos en ACM agrupados por categoría a través del tiempo [5].

2.1.5. *ANÁLISIS CIENCIOMÉTRICO EN EL DESARROLLO DE TEMAS DE TESIS DE MAESTRÍA EN EL PERIODO 1996-2007. CASO DE ESTUDIO CIC*

Tesis publicada en 2010 por Maribel Vázquez [6] en el CIC-IPN, en dónde realiza un análisis de la publicación científica a través de las tesis del instituto en el periodo comprendido de 1996 hasta 2007. Para el análisis se hace uso de indicadores bibliométricos como indicadores de publicación e indicadores de análisis de citas, además de hacer uso del método de palabras asociadas (*co-words)*.

Teniendo esta información describe la producción científica analizando la productividad de los directores de tesis en diversos tópicos de las ciencias de la computación, además se analiza la preferencia o tendencia en determinados tópicos. De igual manera, presenta la información en grafos y diagramas estratégicos con el método de palabras asociadas, como se muestra en Ilustración 2-5.

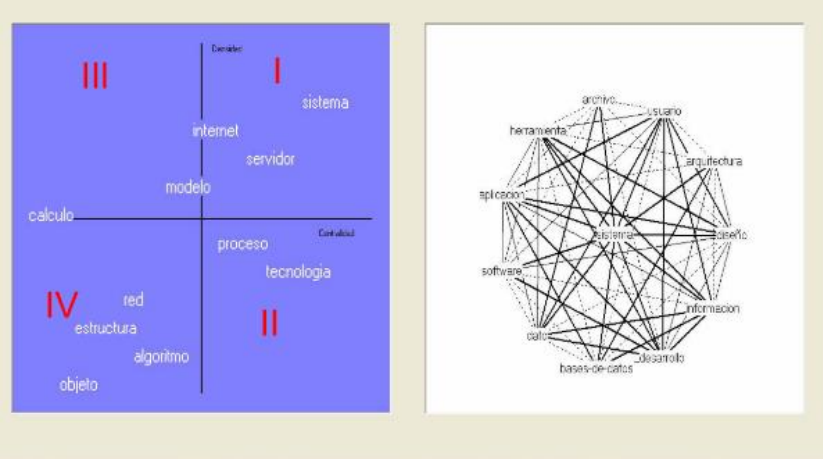


Ilustración 2-5. Diagramas de palabras asociadas en las tesis analizadas [6].

2.1.6. *PROYECTO CYC*

En los años 80, surgieron grandes proyectos que intentaron codificar grandes masas de conocimiento general, uno de los pioneros fue el proyecto "CyC" (actualmente activo) que busca ensamblar una ontología comprensiva y una base de datos de conocimiento general con el fin de permitir a las aplicaciones de inteligencia artificial realizar razonamientos del tipo humano. Su fundador fue Doug Lenat.

El funcionamiento de CyC está basado en reglas simples, por ejemplo: "el agua causa humedad" y la "humedad pudre la comida", permitiendo que el sistema CyC pueda inferir que "el agua pudre la comida". La base de datos contiene aproximadamente 100 000 conceptos y 1 000 000 de declaraciones que abarcan aserciones definidas por humanos, reglas o ideas del sentido común [7].

2.2. COMPARACIÓN CON SISTEMAS DESARROLLADOS BASADOS EN REGLAS

Ambos sistemas desarrollados en el CIC por los doctores Adolfo Guzmán Arenas [8] y Alexander Gelbukh [9].

Tabla 2-1. Comparativa entre sistemas basados en reglas.

	Classifier Demo 3.1	*Clasitex⁺*	*ClassOnto*
Ventana variable acorde a términos relacionados	No	No	Si (Varía el tamaño de la ventana, con respecto a la máxima longitud de los términos relacionados)
Ponderación de segmentos de texto	No	No	Si (Le asigna una importancia a resumen y título)
Multilingüe	Si (inglés, francés, español)	Si (inglés y español)	Si (inglés, español)
Preprocesamiento de texto	Si (Normalización de puntuación y letras capitales)	Si (Normalización de puntuación y letras capitales)	Si (Normalización de puntuación y letras capitales)
Lematización	Si	No	Si
Ámbito de uso	Publicaciones variadas (revistas y periódicos)	Publicaciones variadas (revistas y periódicos)	Publicaciones científicas especializadas (tesis de computación)

2.3. REFERENCIAS DEL CAPÍTULO

[1] E. Spinak, "Indicadores cienciométricos," *ACIMED*, vol. 9, no. SUPPL. 4, pp. 35–41, 2001, [Online]. Available: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352001000400007&lng=es&nrm=iso&tlng=es.

[2] E. Garfield, "Mapping the world of biomedical engineering: Alza lecture (1985)," *Ann. Biomed. Eng.*, vol. 14, no. 2, pp. 97–108, Mar. 1986, doi: 10.1007/BF02584261.

[3] I. Samoylenko, T. C. Chao, W. C. Liu, and C. M. Chen, "Visualizing the scientific world and its evolution," *J. Am. Soc. Inf. Sci. Technol.*, vol. 57, no. 11, pp. 1461–1469, Sep. 2006, doi: 10.1002/asi.20450.

[4] A. R. Figueroa and P. Makagonov, "Modelos de desarrollo del hardware y softare basados en el estudio de computación paralela," *Interciencia*, vol. 32, no. 3, pp. 160–166, 2007, [Online]. Available: http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S0378-18442007000300006.

[5] E. Godínez Fernández, "Desarrollo de un sistema de análisis temático del conocimiento científico," 2009.

[6] M. Vazquez Gallo, "Análisis cienciométrico en el desarrollo de temas de tesis de maestría en el periodo 1996-2007. Caso de estudio: 'Centro de Investigación en Computación del Instituto Politécnico Nacional,'" 2010.

[7] D. B. Lenat and R. Guha, "CYC: A Mid-Term Report," *AI Magazine*, 1990.

[8] A. Guzmán, "Finding the main themes in a spanish document," *Expert Syst. Appl.*, vol. 14, no. 1–2, pp. 139–148, Jan. 1998, doi: 10.1016/S0957-4174(97)00055-9.

[9] A. Gelbukh, "Classifier Demo 3.1." https://nlp.cic.ipn.mx/tools/classifier.

3. Marco teórico

Acrónimos del capítulo

NLP	Natural Langauge Processing (Procesamiento de Lenguaje Natural)
IA	Inteligencia Artificial
TF	Term Frequency (Frecuencia de término)
DF	Document Frequency (Frecuencia de documento)
IDF	Inverse Document Frequency (Frecuencia inversa de documento)
TF-IDF	Term Frequency – Inverse Document Frequency
BoW	Bag of Words (Bolsa de Palabras)
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
ACM	Association for Computing Machinery (Asociación de Maquinaria Computacional)

3.1. INTRODUCCIÓN

En este capítulo se abordarán los fundamentos teóricos principales de la cienciometría, las ciencias relacionadas, así como del modelado de tópicos, y la representación y razonamiento del conocimiento.

3.2. CIENCIOMETRÍA

La cienciometría es una disciplina que aporta los instrumentos e indicadores básicos necesarios para evaluar de forma objetiva la estructura, dinámica, producción y evolución de los componentes que integran y relacionan las distintas disciplinas y áreas de un campo de conocimiento científico.

Estos indicadores son utilizados para el fortalecimiento de líneas de investigación en distintos campos, porque proporcionan información estratégica relacionada con las tendencias, problemas, enfoques y temas de investigación desarrollados. De igual manera, estos indicadores ayudan en la tarea de descripción de las características, estructura y dinámicas de la producción científica, elementos fundamentales para trazar nuevos procesos y estrategias de formación e investigación disciplinar, en este caso, mediante el uso de técnicas especializadas que ayudan a identificar patrones de conocimiento que subyacen en la red de interacciones entre el conjunto de actores y el conjunto de trabajos publicados [1].

La cienciometría parte de la base de que los resultados de las investigaciones científicas y técnicas se plasman en forma escrita a través de cinco dimensiones principales, también llamada rosa de los vientos de la investigación Ilustración 3-1. Estas dimensiones son tomadas en consideración por la cienciometría para la evaluación del desarrollo de la ciencia [2].

- Comunidad científica
- Políticas públicas
- Sistema de enseñanza
- Mercado
- Administración y divulgación

Ilustración 3-1. Rosa de los vientos de la investigación [2].

3.2.1. *BIBLIOMETRÍA*

La bibliometría es un campo de la cienciometría que estudia la naturaleza y el curso de una disciplina mediante métodos estadísticos y matemáticos de la publicación científica escrita. Para ello se ayuda de indicadores bibliométricos, que proporcionan información sobre los resultados de la actividad científica en cualquiera de sus manifestaciones.

3.2.1.1. INDICADORES BIBLIOMÉTRICOS

Se muestra los principales indicadores bibliométricos en la Tabla 3-1. Los cuales proporcionan información específica sobre el volumen y el impacto de las actividades de investigación científicas.

Tabla 3-1. Principales indicadores bibliométricos.

Indicadores bibliométricos
Indicadores de Actividad Científica
• Número y distribución de publicaciones (artículos, patentes, libros, etc.) • Publicaciones producidas • Productividad
Indicadores de Impacto
• Número de citas recibidas • Medición del factor de impacto

3.2.2. *DIFERENCIA ENTRE BIBLIOMETRÍA Y CIENCIOMETRÍA*

La bibliometría se superpone considerablemente con la cienciometría [3]. En realidad, ambas ciencias no son las mismas. La bibliometría es un subconjunto de la cienciometría, y se limita al análisis de publicaciones y sus propiedades [4]. Pero todos se trata de métricas y aspectos de medición de la investigación científica, el mapeo de disciplinas y los resultados de la investigación. En su tipología para la definición y clasificación de estas disciplinas, McGrath [5] identificó su objeto de estudio, sus variables, sus métodos y sus objetivos, como se muestra en la Tabla 3-2.

Tabla 3-2. Tipología entre bibliometría y cienciometría.

Tipología	*Bibliometría*	*Cienciometría*
Objeto de estudio	Libros, documentos, revistas, artículos, autores y usuarios.	Disciplinas, materias, campos, esferas.
Variables	Números en circulación, citas, frecuencia de aparición de palabras, etc.	Aspectos que diferencias a las disciplinas y a las subdisciplinas.
Métodos	Clasificación, frecuencia, distribución.	Análisis de conjunto y de correspondencia.
Objetivos	Asignar recursos, tiempo, dinero, etc.	Identificar esferas de interés, comunicaciones entre científicos, etc.

3.2.3. *APLICACIÓN DE LA CIENCIOMETRÍA Y LA BIBLIOMETRÍA*

La importancia de las técnicas bibliométricas y cienciométricas puede notarse al analizar la siguiente lista de posibilidades de aplicación [6]:

- Identificar las tendencias y el crecimiento del conocimiento en las distintas disciplinas.
- Identificar autores y tendencias en distintas disciplinas.
- Medir la utilidad de los servicios de diseminación selectiva de la información.
- Predecir tendencias en la publicación.
- Identificar las entidades núcleo de cada disciplina.
- Estudiar la dispersión y la obsolescencia de la literatura científica.
- Diseñar normas para estandarización.
- Diseñar procesos automáticos de indización, clasificación y confección de resúmenes.
- Predecir la productividad de editores, autores, entidades, etc.

3.3. LENGUAJE NATURAL Y PROCESAMIENTO DE LENGUAJE NATURAL

Los lenguajes naturales son aquellos que están involucrados y son utilizados de forma natural por los seres humanos, para propósitos de comunicación. El procesamiento de lenguaje natural (NLP, por sus siglas en inglés) es una rama de la inteligencia artificial, concretamente del aprendizaje máquina, que ayuda a las computadoras a entender, interpretar y manipular el lenguaje humano [7].

Las técnicas de NLP se desarrollaron con la finalidad de que las computadoras comprendieran los comandos establecidos en el lenguaje natural, y responder acorde a tal solicitud. Junto con estas técnicas, NLP es un área interdisciplinaria (Ilustración 3-2), que comprende varias técnicas y herramientas.

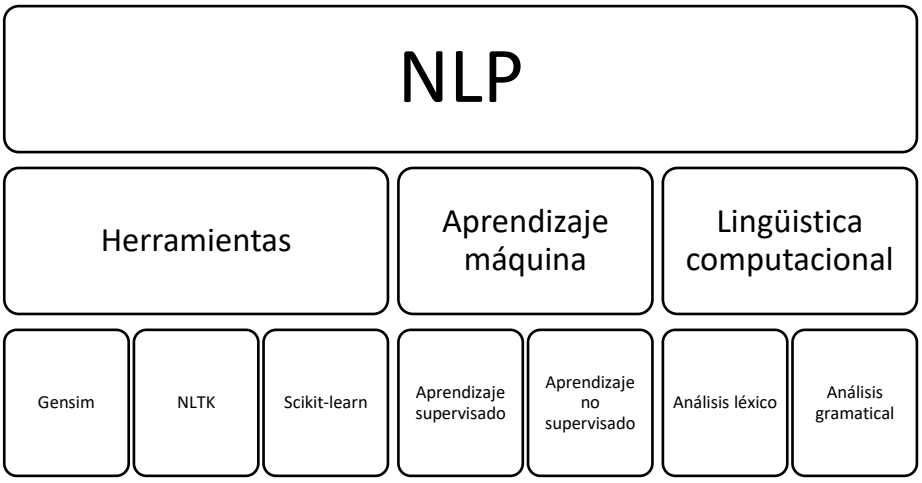


Ilustración 3-2. NLP como área interdisciplinaria, principales áreas y herramientas.

3.3.1. *LINGÜÍSTICA COMPUTACIONAL*

La lingüística computacional estudia cómo construir los modelos del lenguaje de tal manera que sean entendibles para las computadoras, eso quiere decir que no solo se analiza el uso del lenguaje en el comportamiento humano, sino también se aplican métodos formales que permitan la formulación exacta de las hipótesis y su posterior verificación automática utilizando los datos lingüísticos (los corpus) [8].

3.3.2. *MINERÍA DE TEXTOS*

La minería de texto es el proceso de obtener información de alta calidad del texto. Es un campo multidisciplinario que puede rastrear sus raíces a la teoría y la práctica de la minería de datos, también llamada "descubrimiento de conocimiento en bases de

datos" [9]. Por lo general, implica el proceso de obtener el corpus de textos, estructurar el corpus, limpiarlos, encontrar patrones dentro de los datos estructurados y, finalmente, evaluar e interpretar la salida; y con esta identificación de patrones ocultos en grande corpus de documentos de texto no estructurados, se busca lograr descubrir conocimiento que era incluso desconocido para los autores de los documentos, vale la pena mencionar que hace falta el análisis del usuario para deducir el nuevo conocimiento.

Entre sus principales aplicaciones se encuentra:

- Apoyar en las investigaciones.
- Agrupamiento de textos que tienen mayor semejanza.
- Categorización por tópicos o temas.
- Facilitar la recuperación, análisis y visualización de la información.
- Navegación en los corpus de documentos.

3.3.3. *MODELADO DE TÓPICOS*

El modelado de tópicos o temas es una tarea de procesamiento de lenguaje natural, que debe encontrar los temas que el documento analizado. En aprendizaje automático y el procesamiento del lenguaje natural, los modelos de temas o tópicos son modelos generativos, que proporcionan un marco probabilístico [10] para analizar un conjunto de documentos, detectar patrones de palabras y frases dentro de ellos, y agrupar automáticamente palabras y expresiones similares que mejor caracterizan un conjunto de documentos [11].

Los "temas" significan las variables ocultas, a ser estimados, que vinculan las palabras en un vocabulario y su aparición en documentos. Un documento se ve como una mezcla de temas. Los modelos de temas descubren los temas ocultos en toda la colección y registran los documentos analizados de acuerdo con esos temas, cada palabra sería una extracción de los temas descubiertos. En otras palabras, un documento se considera una mezcla de un conjunto de temas; y un tema es una distribución sobre un vocabulario fijo, y estos temas se generan a partir de la colección de documentos [12]. Por último, se genera un documento con la distribución de sus temas.

Si se realiza esta tarea con un conjunto de documentos que no está previamente etiquetados, se clasificaría como como aprendizaje automático "no supervisado" porque no requiere una lista predefinida de etiquetas o datos de entrenamiento que hayan sido clasificados previamente por humanos. De lo contrario sería aprendizaje automático "supervisado".

3.3.3.1. PALABRAS CO-OCURRENTES

El método o análisis de palabras co-ocurrentes o palabras asociadas se usa para tareas de modelado de tópicos y consiste en la detección de palabras que caracterizan un tema y en contar la co-aparición de estas, en campos tales como títulos de artículos, palabras clave, resúmenes o bien directamente en el texto libre.

3.3.4. *PROCESAMIENTO DE TEXTOS*

En el área de procesamiento de lenguaje natural y la computación, para que una máquina sea capaz de analizar un texto o conjuntos de textos (corpus), debe primero realizarse una serie de acciones; que dependen en su implementación, de cómo se llevará a cabo el análisis. A continuación, se mencionan las más relevantes.

3.3.4.1. FRECUENCIA DE TÉRMINO (TF)

Para conocer la importancia de una palabra dentro de un documento se puede utilizar la frecuencia de palabras como medida.

Siendo:

$$tf_{w,d} = \frac{w}{d}$$

Donde:

- w es una palabra
- d es el número total de palabras del documento
- $tf_{w,d}$ es el número de veces que se repite w en d, y siempre será un valor positivo y real.

Sin embargo, no todas las palabras en un documento son igualmente importantes, es decir, algunas pueden aportar más información que otras con respecto al documento, y dado que $tf_{w,d}$ considera todas las palabras igual de importantes es una desventaja.

3.3.4.2. FRECUENCIA INVERSA DE DOCUMENTO (IDF)

Al tratar con corpus de texto, se puede manejar la importancia de una palabra dependiendo de su frecuencia de aparición en la colección de documentos.

Siendo:

$$df_t = \frac{d_w}{d_c}$$

Donde:

- d_c es el número total de documentos en el corpus
- d_w es el número de documentos con la palabra
- df_t es el número de documentos d_w que contienen w en un corpus c

Siendo que si w aparece en un gran número de documentos, su df_t será alto, lo que significa que el término se encuentra en una gran cantidad de documentos, por el contrario, si el valor es bajo, aparece en pocos documentos. Dado que se buscar resaltar aquellas palabras que mejor definan a un documento, entonces las palabras "únicas" deberían tener un gran valor, por lo que se plantea el idf_t.

Siendo:

$$idf_t = \log(df_t^{-1})$$

Utilizando un logaritmo para amortiguar la importancia un término que tenga alta frecuencia, por ejemplo: usando logaritmo base 2, si se tiene 1 000 000 en df_t^{-1}, se amortiguaría a 19.9 el valor de idf_t

3.3.4.3. MEDIDA TF-IDF

La medida TF-IDF combina la medida TF e IDF, dando como resultado un valor que corresponde a cada término en cada documento, indicando el grado de relevancia de una palabra para un documento con respecto a un corpus; y su fórmula es la siguiente:

$$tf - idf = tf_{w,d} * idf_t$$

El valor de esta medida es:

- Alto: cuando el término se encuentra muchas veces en un número pequeño de documentos, indicando así que el término es importante.
- Bajo: cuando el término aparece pocas veces en un documento o en muchos documentos, indicando así que el término no es importante.

Una de las desventajas de usar TF-IDF es que no toma en consideración el significado semántico, ya que considera la importancia de las palabras debido a como pesan, y esto no significa que tome en cuenta el contexto de las palabras, ni la importancia de su uso en ese contexto. También al igual que BoW (Bag of Words, bolsa de palabras

en español), TF-IDF ignora el orden de las palabras, y de ahí que sustantivos complejos como "Benemérito de las Américas", no sea considerado como una unidad o "concepto".

3.3.4.4. N-GRAMAS

Un n-grama es una subsecuencia de n elementos de una secuencia dada. Así se puede aplicar para el procesamiento de lenguaje natural, al tener la necesidad de involucrar mayor información analizando el contexto de los elementos en una secuencia. Los n-gramas pueden tener como elementos:

- Elementos léxicos (palabras, lemas o raíces)
- Etiquetas de las categorías gramaticales
- Nombre de las relaciones semánticas
- Caracteres
- N-gramas sintácticos mixtos (combinaciones de los tipos anteriores)

3.3.4.5. LDA

Es una técnica de procesamiento de lenguaje natural que permite explicar conjuntos de observaciones mediante grupos no observados. Aplicando al modelado de tópicos, plantea que, si las observaciones son palabras reunidas en documentos, entonces cada documento es una mezcla de una pequeña cantidad de temas y que la presencia de cada palabra es atribuible a uno de los temas del documento. Por lo tanto, si una colección de documentos es lo suficientemente grande, LDA descubrirá dichos conjuntos de términos (temas) basándose en la co-ocurrencia de términos individuales. Aunque la asignación de una etiqueta significativa a un tema individual depende del usuario.

3.3.4.6. LSA

Es una técnica de procesamiento de lenguaje natural que analiza las relaciones entre un conjunto de documentos y los términos que contienen. Lo realiza asumiendo que las palabras que tienen un significado cercano aparecerán en fragmentos de texto similares; utilizando grandes cantidades de documentos. LSA no hace uso de analizadores sintácticos, morfológicos, o relaciones semánticas, así como tampoco de recursos construidos manualmente como diccionarios o tesauros.

3.3.4.7. MODELO ESPACIO VECTORIAL

Es un modelo ampliamente usado en las ciencias de la computación. Su amplio uso se debe a la simplicidad del modelo y de su muy clara base conceptual que corresponde a la intuición humana en el procesamiento de información y de datos [8].

Consiste en describir objetos utilizando una representación con sus rasgos (características) y sus valores. En el ámbito del NLP, el modelo espacio vectorial de un documento sería representarlo a través las palabras que incluya. El modelo ya construido tendría un espacio de *N* dimensiones, y cada dimensión en este espacio corresponde a una de las características (en éste caso serían cada palabra, frase, etc.): siendo el número de dimensiones, igual al número de las características que tiene el objeto en nuestro modelo [8].

3.3.4.8. TOKENIZACIÓN

Este proceso consiste en dividir las cadenas de texto en características que contengan algún significado para el procesamiento, ya sea palabras, frases o párrafos. Con esto se omitirían los signos de puntuación, espacios en blanco, así como el uso de mayúsculas y minúsculas.

3.3.4.9. ELIMINACIÓN DE "*STOPWORDS*"

La eliminación de "*stopwords*" o palabras vacías, consiste en identificar un conjunto de palabras que aporta muy poco o ningún significado semántico en los textos, generalmente son palabras que aparecen con mayor frecuencia en un idioma, y suelen ser las preposiciones, pronombres, verbos auxiliares, artículos, etc. [13].

3.3.4.10. LEMATIZACIÓN

Es un proceso que consiste en dad una forma flexionada, (es decir un plural, un femenino, conjugada, etc.) hallar el lema correspondiente. Siendo un lema la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra. El uso de este proceso está determinado por la aplicación del análisis, siendo que en algunos análisis no es requerido.

3.3.4.11. ELIMINACIÓN DE SINÓNIMOS

Este proceso está sujeto a la aplicación del análisis, de igual manera que la lematización. Consiste en obtener conjuntos de sinónimos de las palabras, y llegar a un consenso en la utilización de estas palabras, para que no haya "conceptos" duplicados.

3.4. REPRESENTACIÓN Y RAZONAMIENTO DEL CONOCIMIENTO

La representación del conocimiento es el campo de la inteligencia artificial (IA, por sus siglas en inglés) enfocado a representar la información de una forma que un sistema informático pueda utilizar para realizar actividades o resolver tareas complejas; buscando como los humanos resuelven y estructuran el conocimiento, y tratando de imitarlo.

Las características que debe cubrir un sistema de representación del conocimiento son las siguientes:

- Cobertura: la representación del conocimiento debe cubrir la información en anchura y profundidad. Sin una cobertura amplia, la representación del conocimiento no puede determinar nada ni resolver ambigüedades.
- Comprensible por humanos: la representación del conocimiento es vista como un lenguaje natural, así que la lógica debería fluir libremente.
- Consistencia: la representación del conocimiento debe eliminar conocimiento redundante o conflictivo.
- Facilidad de modificación y actualización.

Algunos formalismos de representación del conocimiento son:

- Recursos no ontológicos (glosarios, léxicos, esquemas de clasificación, tesauros, etc.)
- Ontologías
- Redes semánticas: árboles de conocimientos.

Para la investigación se ocupa el sistema de clasificación de computación ACM, el cual se define como: "ontología poli jerárquica".

3.4.1. *ONTOLOGÍAS*

En las ciencias de la computación, una ontología es una representación del conocimiento, que puede estar limitada a una disciplina o campo académico, con el fin de limitar la complejidad y organizar los datos en información y conocimiento, mostrando las propiedades de una o varias áreas temáticas y cómo se relacionan, mediante la definición de un conjunto de conceptos y categorías.

3.4.2. *ÁRBOL DE CONOCIMIENTOS*

Un árbol de conocimientos se define como una red semántica jerárquica que representa el conocimiento en forma de una red de árbol o una estructura de árbol; teniendo un valor en la raíz y subárboles en un nodo padre. Cada nodo puede representar un concepto, un conjunto de conceptos o una temática. Y éstos a su vez pueden estar definidos con términos.

Definiéndose como un término a un símbolo convencional para un concepto, poniendo de manifiesto que el término puede ser tanto una unidad léxica simple como la combinación de varias unidades.

3.5. REFERENCIAS DEL CAPÍTULO

[1] J. H. Ávila Toscano, *Cienciometría y bibliometría. El estudio de la producción científica. Métodos, enfoques y aplicaciones en el estudio de las Ciencias Sociales*. 2018.

[2] P. Escorsa and R. Maspons, "De la vigilancia tecnológica a la inteligencia competitiva," *Revista Madrid*, Madrid, pp. 1–20, May 01, 2001.

[3] M. Thelwall, "Bibliometrics to webometrics," *J. Inf. Sci.*, vol. 34, no. 4, pp. 605–621, Aug. 2008, doi: 10.1177/0165551507087238.

[4] G. Y. (Université du Q. à Montréal), *Bibliometrics and Research Evaluation: Uses and Abuses (History and Foundations of Information Science)*. The MIT Press, 2016.

[5] W. McGrath, "What bibliometricians, scientometricians and informetricians study; a typology for definition and classification; topics for discussion," 1989.

[6] E. Spinak, "Indicadores cienciométricos," *ACIMED*, vol. 9, no. SUPPL. 4, pp. 35–41, 2001, [Online]. Available: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352001000400007&lng=es&nrm=iso&tlng=es.

[7] "Qué es el Procesamiento de Lenguaje Natural - Natural Language Processing? | SAS." https://www.sas.com/es_ar/insights/analytics/what-is-natural-language-processing-nlp.html (accessed Jun. 18, 2021).

[8] G. Sidorov, *Construcción no-lineal de n-gramas en la lingüistica computacional*. 2013.

[9] M. Ponweiser, "Latent Dirichlet Allocation in R," Vienna University of Business and Economics, 2012.

[10] Z. Tong and Z. Haiyi, "A Text Mining Research Based on LDA Topic Modelling," *Comput. Sci. Inf. Technol.*, pp. 201–210, 2016, doi: 10.5121/csit.2016.60616.

[11] B. Grun and K. Hornik, "topicmodels: An R Package for Fitting Topic Model," *J. Stat. Softw.*, vol. 40, no. 13, 2011.

[12] D. M. Blei, "Probabilistic Topic Models," *Commun. ACM*, vol. 55, no. 4, pp.

77–84, 2012.

[13] B. Patel and D. Shah, “Significance of stop word elimination in meta search engine,” in *2013 International Conference on Intelligent Systems and Signal Processing, ISSP 2013*, 2013, pp. 52–55, doi: 10.1109/ISSP.2013.6526873.

4. Análisis y Diseño

Acrónimos del capítulo

UML	Unified Modeling Language (Lenguaje de Modelado Unificado)
ACM	Association for Computing Machinery (Asociación de Maquinaria Computacional)
PDF	Portable Document File (Formato de archivo "Formato de Documento Portátil")
TXT	Formato de archivo que contiene texto plano
DAT	Formato de archivo que puede contener cualquier información
CSV	Comma Separated Values (Formato de archivo "Valores Separados por Comas")
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
TF-IDF	Term frequency – Inverse document frequency (Frecuencia de término – Frecuencia inversa de documento)
F1	Una medida que combina precisión y recuperación es la media armónica de precisión y recuperación.

4.1. INTRODUCCIÓN

En la etapa de análisis se tiene como objetivo estudiar y comprender el dominio del problema, lo cual se puede resumir a responder la interrogante ¿qué hacer? [1].

En la etapa de diseño de un nuevo sistema, se definen los objetivos de diseño, seleccionando formas y estrategias en la construcción, y estableciendo requerimientos de plataforma (hardware y software) respondiendo a la cuestión ¿cómo hacerlo? [2].

4.2. DESCRIPCIÓN DEL SISTEMA

Como se describió en la propuesta de solución, el sistema final tendrá cinco bloques principales que está basada en la metodología utilizada para la realización de análisis cienciométricos de Michán y Muñoz [3].

- Recuperación
- Migración
- Análisis
- Visualización
- Interpretación

Para poder categorizar las publicaciones, se necesitará realizar:

1. Se parte de una ontología, en donde están las clasificaciones que buscamos asociar a las publicaciones.
2. Se crea un árbol de conocimientos basándose en dicha ontología, en donde cada clasificación del ACM tenga términos que la definan.

El sistema CLASSONTO que se implementará para el análisis consistirá en que sea capaz de etiquetar la publicación bajo un árbol de conocimientos basado en la ontología definida por ACM para las ciencias de la computación.

4.2.1. *ALGORITMO DE CLASIFICACIÓN DE CLASSONTO*

El algoritmo de clasificación de CLASSONTO se describe de la manera siguiente:

Input:

text (Str): Texto a analizar

terms (Dict): Árbol de conocimiento, como base para las votaciones

vote (Bool): Selección para "propagación de votos"

worth (Int): Importancia que tiene el texto a analizar

Output:

Archivo DAT conteniendo las clasificaciones encontradas con sus respectivos votos

Archivo DAT conteniendo los términos encontrados, su frecuencia de aparición y las clasificaciones que representa

Archivo DAT conteniendo las palabras que no representan clasificación, con su respectiva frecuencia de aparición

Variables:

len_max Tamaño máximo de la ventana, determinado por el valor máximo que tiene los términos en el árbol de conocimiento

len_win Tamaño actual de la ventana

window Contenido de la ventana

index Puntero inicio de la ventana

---

1. *index = Str[begin]*
2. Analizar *window*
3. While *index* != Str[end]
   - 3.1. If (*window* exist in *Dict*):
     - 3.1.1. Votación(*window, vote, worth*)
     - 3.1.2. *index += len_win*
     - 3.1.3. go to step 2
   - 3.2. Else:
     - 3.2.1. If l*en_win* > 1:
       - 3.2.1.1. *len_win* -= 1
       - 3.2.1.2. go to step 2
     - 3.2.2. Else:
       - 3.2.2.1. NoEncontradas(*window*)
       - 3.2.2.2. *index += len_win*
       - 3.2.2.3. *len_win = len_max*
       - 3.2.2.4. go to step 2

4.2.2. *ÁRBOL DE CONOCIMIENTOS*

El árbol de conocimientos del sistema CLASSONTO está basado en la ontología multi-jerárquica de ACM para la computación. A partir de ahí se ha enriquecido el árbol con términos relacionados con clasificaciones tomando como base la investigación de dichas clasificaciones (páginas web, artículos, libros) y por último con la terminología que está presente en las tesis analizadas; esto para ambos idiomas (español e inglés).

La sugerencia de términos para el enriquecimiento del árbol de conocimiento es a través de su frecuencia de aparición (términos no reconocidos por el sistema), tanto de manera individual como colectiva (en todo el corpus de tesis analizadas). A su vez este conjunto de términos no reconocidos se le realiza una poda, para todavía reducir más el espacio de búsqueda de términos que representen una clasificación.

Dado que un árbol de conocimientos contiene un número finito de conceptos o categorías en un patrón estructurado a las que se busca acotar al momento de clasificar las tesis, se descartarían el uso de modelos entrenados, LDA y LSA.

4.2.3. *PODA DE TÉRMINOS NO RECONOCIDOS EN EL CORPUS*

TF-IDF sería una elección razonable para el filtrado de términos en una colección de documentos debido a que es una medida numérica que expresa cuán relevante es una palabra para un documento en una colección, sin embargo, para los fines del sistema no es válido, debido a que en TF-IDF las palabras más comunes en una colección de documentos tendrán menor valor que las que ocurren ocasionalmente, y esto no es aplicable porque el sistema se basa en un árbol de conocimiento obtenido a partir de una ontología, y esta ontología esta desarrollada a partir de un consenso de términos que denotan temáticas. De ahí que estos términos puedan ser repetitivos en el corpus de documentos, y brindar información a la hora de clasificar por temáticas.

Por lo que la reducción de términos no reconocidos es de la siguiente forma:

- Eliminación de términos que solamente se encuentran en una tesis.
- Eliminación de términos que tienen menor frecuencia que un límite (establecido en 10) tomando en cuenta también su aparición en tesis, es decir si en 4 tesis, existe el conteo de 40 veces que aparece un término, descartarlo.

4.2.4. *PONDERACIÓN DE TEXTO*

CLASSONTO permite ponderar o darle una importancia al texto que analiza, permitiendo asignarles valores a los principales elementos del documento de tesis:

- Título
- Resumen
- Texto completo

La asignación de valores se basa con respecto a que tanta información debería brindar cada elemento, conforme a la clasificación de la tesis, siendo el título el que mayor valor tendrá, seguido del resumen y por último el texto completo.

Los valores usados en esta investigación son:

- Título: 50
- Resumen: 15
- Texto completo: 1

Los valores mostrados representan el factor por el cual se va a multiplicar una vez se emita un voto, es decir si se produce una votación de una clasificación, ésta se multiplicará por su correspondiente factor.

Estos valores fueron propuestos debido a que, sin ponderación, las clasificaciones más importantes pueden superar en promedio más de 150 en sus votos. Teniendo esto en cuenta se considera que una relevancia del título de 1/3 del promedio de los votos es adecuada. Para el resumen, se tiene el problema que es muy variante en su tamaño (desde ¼ de página hasta 2 páginas); debido a esto y a que un resumen puede contener varias clasificaciones que sean votadas, se propone que el valor del resumen sea del 30% con respecto al título.

4.2.5. *VENTANEO*

Al analizar un texto, se busca tener el contexto en el que es empleado las palabras, siendo útil las técnicas de utilizar n-gramas, donde se propone de 4 a 6 n-gramas de palabras (Google usa para una aplicación n-gramas de tamaño 5, http://storage.googleapis.com/books/ngrams/books/datasetsv2.html) para poder decir que se ha tomado en cuenta el contexto. CLASSONTO aborda este problema utilizando una ventana, teniendo como característica fundamental el manejo de una ventana de tamaño variable, debido a que su tamaño está supeditado a la máxima longitud de la terminología de las clasificaciones, es decir, si en el árbol de conocimiento hay una frase compuesta por cinco palabras que relaciona a una clasificación, y esa es la frase relacionada con la mayor longitud, entonces la ventana del algoritmo sería de tamaño cinco, permitiendo lidiar con acrónimos y/o expresiones en el ámbito de la computación, para la investigación se obtuvo un máximo de tamaño de ventana de valor 6.

4.2.6. *PREPROCESAMIENTO*

El preprocesamiento del texto se realiza con la herramienta spaCy para limpiar y simplificar el texto, siguiendo un proceso o "*pipeline*" en el cual se eliminan caracteres indeseados (paréntesis, signos, etc.), cadenas de e-mails y urls, así como también se normaliza el texto, eliminando los *stopwords* y palabras selectas por sus *POS tags* (determinantes y adposiciones) y por último se cambia el texto a sus lemas.

4.2.7. *PROPAGACIÓN DE VOTO*

En el árbol de conocimiento hay clasificaciones (nodos hijo) que pueden ser parte de una clasificación más general (nodo padre), por ejemplo "Aprendizaje de máquina" y "Aprendizaje supervisado" siendo este último el nodo hijo del primero; por lo que CLASSONTO al momento de incrementar el contador de una clasificación, podrá transmitir ese "voto" de una clasificación de jerarquía de nivel bajo a una clasificación de jerarquía de nivel alto.

La propagación se hace a través de un factor, para esta investigación se utiliza ¼ y conforme se propaga hacia las clasificaciones más generales (nodos padres), este factor va aumentando exponencialmente. Esta configuración permite que las clasificaciones más generales no se vean beneficiadas de gran manera gracias a las clasificaciones específicas.

Un ejemplo para esta forma de propagación es la siguiente:

Se realiza una votación para "*11.4.1.1. Supervised learning*" que es clasificación de nivel 4 (tiene 3 nodos padres) y que tiene forma parte de una clasificación más general "*11. Computing methodologies*" que es nivel 1. Sus factores de propagación serían los siguientes:

"*11.4.1.1. Supervised learning*": $\left(\frac{1}{4}\right)^0 = 1$

"*11.4.1. Learning paradigms*": $\left(\frac{1}{4}\right)^1 = \frac{1}{4}$

"*11.4. Machine learning*": $\left(\frac{1}{4}\right)^2 = \frac{1}{16}$

"*11. Computing methodologies*": $\left(\frac{1}{4}\right)^3 = \frac{1}{64}$

Se puede observar que entre más específica sea la clasificación que reciba el voto, menor será su impacto en la clasificación general. Evidentemente esto suma con las

demás clasificaciones obtenidas, con la posibilidad de que varias clasificaciones específicas distintas beneficien a una clasificación general.

4.2.8. *ASIGNACIÓN DE "IMPORTANCIA" A CLASIFICACIONES*

La asignación de las etiquetas "importancia" en las clasificaciones (alta, media o baja) es por medio de una proporción con respecto a la clasificación que fue mayormente votada.

Tomando como máximo el valor de la clasificación con el mayor puntuaje de votación, se tienen tres intervalos que determinarán el valor de la etiqueta "importancia".

Si los votos de la clasificación son:

- Menor a $\frac{1}{3}$ del valor máximo de votación → etiqueta: BAJA
- Entre $\frac{1}{3}$ del valor máximo de votación y $\frac{2}{3}$ → etiqueta: MEDIA
- Mayor a $\frac{2}{3}$ del valor máximo de votación → etiqueta: ALTA

4.2.9. *PODA DE CLASIFICACIONES PREDICHAS*

El sistema CLASSONTO al reconocer términos, genera votaciones para todas las clasificaciones que están relacionadas, sin embargo, esto genera que términos con baja frecuencia de aparición también sean votados, y por ende que aparezcan muchas clasificaciones; por tal motivo, surge la necesidad de realizar una poda de clasificaciones encontradas. Para esta investigación se realizó una poda dejando solo las 10 clasificaciones con mayor votación. El valor es debido al rango recomendado de *keywords* utilizado para la descripción un trabajo, que usualmente va de 3 a 7 *keywords*. Por lo que, partiendo de la cota inferior, y considerando que las clasificaciones predichas tendrán un nivel de importancia (Alto, Medio, Bajo), se propone el valor de 10 clasificaciones para describir el trabajo.

4.2.10. *VALIDACIÓN DE RESULTADOS*

La validación de resultados se obtiene con un reporte de clasificación *F1*, *precision*, *recall* y *support* (F1, precisión, recuperación y soporte) y matrices de confusión, que comparan las clasificaciones obtenidas por CLASSONTO contra el *Golden Standard*. De igual manera se proponen tres modos de evaluación tomando en cuenta que se está usando una clasificación multi-jerárquica y por ende multi-etiqueta y multi-clase. Siendo los modos de evaluación propuestos:

Estricto: Penaliza una clasificación encontrada que no coincide con la etiqueta (importancia) manejada por el Golden Standard.

- Con coincidencia perfecta, otorga calificación de 10.
- Con coincidencia, otorga calificación de 8.5.
- Sin coincidencia no otorga calificación.
- Calcula el resultado dividiendo con todas las posibles coincidencias perfectas.

Bondadoso: Otorga un bonus al encontrar una clasificación que coincide con la etiqueta (importancia) manejada por el Golden Standard.

- Con coincidencia perfecta, otorga calificación de 15 (10 más bonus de 5).
- Con coincidencia, otorga calificación de 10.
- Sin coincidencia no otorga calificación.
- Calcula el resultado dividiendo con todas las posibles coincidencias.

Onehot: No considera las etiquetas (importancia), y realiza la evaluación solamente tomando en cuenta la identificación de las clases.

- Con coincidencia, otorga calificación de 1.
- Sin coincidencia no otorga calificación.
- Calcula el resultado dividiendo entre todas las posibles coincidencias.

Los modos propuestos tienen como finalidad poder detectar cuando el sistema CLASSONTO identificó los temas y les asignó las etiquetas de "importancia" correctamente; caso contrario a los modos estándar *F1*, *precision*, *recall* y *support* que no permiten resaltar esta situación.

4.2.11. *VISUALIZACIÓN Y EVALUACIÓN*

La evaluación es de manera temporal, es decir, se lleva a cabo con las clasificaciones encontradas en las tesis, y tomando como parámetro el año que dichas tesis fueron publicadas.

La visualización de los resultados es por medio de tablas y gráficas, que se obtengan a partir de la evaluación temporal y de los resultados de clasificación del sistema.

4.2.12. *INFERENCIA DE CLASIFICACIONES EN TERMINOLOGÍA NO RECONOCIDA*

La inferencia de clasificaciones permite reducir el listado de palabras no reconocidas por el sistema CLASSONTO y sugerir clasificaciones que las relacionen. El proceso

se realiza teniendo en cuenta si las palabras desconocidas se presentan en tesis clasificadas similarmente. Se tiene la posibilidad de seleccionar el margen con respecto a la clasificación, es decir, se puede acotar a solamente sugerir clasificaciones teniendo como base solamente las clasificaciones de importancia "Alta" predichas.

4.2.13. *FUNCIONAMIENTO DE ALGORITMO CLASSONTO*

A continuación, se muestra un ejemplo de su funcionamiento en donde:

- El texto remarcado en amarillo significa la ventana.
- El texto remarcado en verde significa que la ventana encontró una coincidencia.
- El valor remarcado en rojo significa el valor pasado (votación) que tenía la clasificación.
- El valor remarcado en azul significa el valor nuevo (votación) que tiene la clasificación.
- Las letras en cursiva significan la salida del sistema.

Se empieza con una ventana de tamaño establecido, para este ejemplo es de tamaño 4, que se va recorriendo a través del texto. De la Ilustración 4-1 hasta la Ilustración 4-8 se muestra el recorrido de la ventana.

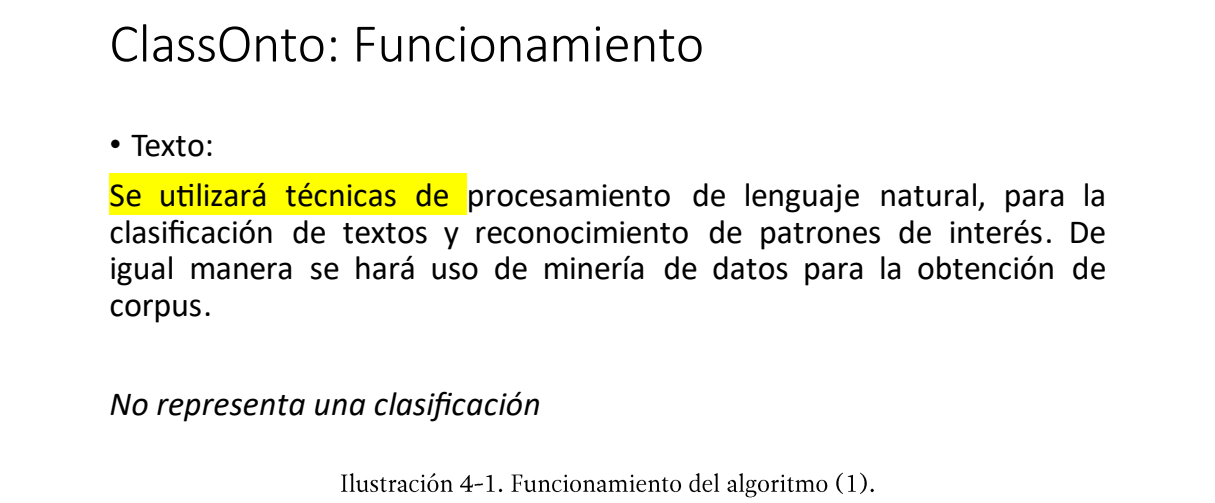


Ilustración 4-1. Funcionamiento del algoritmo (1).

ClassOnto: Funcionamiento

• Texto:

Se utilizará técnicas de procesamiento de lenguaje natural, para la clasificación de textos y reconocimiento de patrones de interés. De igual manera se hará uso de minería de datos para la obtención de corpus.

No representa una clasificación

Ilustración 4-2. Funcionamiento del algoritmo (2).

ClassOnto: Funcionamiento

• Texto:

Se utilizará técnicas de procesamiento de lenguaje natural, para la clasificación de textos y reconocimiento de patrones de interés. De igual manera se hará uso de minería de datos para la obtención de corpus.

No representa una clasificación

Ilustración 4-3. Funcionamiento del algoritmo (3).

ClassOnto: Funcionamiento

• Texto:

Se utilizará técnicas de procesamiento de lenguaje natural, para la clasificación de textos y reconocimiento de patrones de interés. De igual manera se hará uso de minería de datos para la obtención de corpus.

No representa una clasificación

Ilustración 4-4. Funcionamiento del algoritmo (4).

ClassOnto: Funcionamiento

• Texto:

Se utilizará técnicas de procesamiento de lenguaje natural, para la clasificación de textos y reconocimiento de patrones de interés. De igual manera se hará uso de minería de datos para la obtención de corpus.

No representa una clasificación

Ilustración 4-5. Funcionamiento del algoritmo (5).

ClassOnto: Funcionamiento

• Texto:

Se utilizará técnicas de procesamiento de lenguaje natural, para la clasificación de textos y reconocimiento de patrones de interés. De igual manera se hará uso de minería de datos para la obtención de corpus.

No representa una clasificación

Ilustración 4-6. Funcionamiento del algoritmo (6).

ClassOnto: Funcionamiento

• Texto:

Se utilizará técnicas de procesamiento de lenguaje natural, para la clasificación de textos y reconocimiento de patrones de interés. De igual manera se hará uso de minería de datos para la obtención de corpus.

No representa una clasificación

Ilustración 4-7. Funcionamiento del algoritmo (7).

ClassOnto: Funcionamiento

• Texto:

Se utilizará técnicas de procesamiento de lenguaje natural, para la clasificación de textos y reconocimiento de patrones de interés. De igual manera se hará uso de minería de datos para la obtención de corpus.

No representa una clasificación

Ilustración 4-8. Funcionamiento del algoritmo (8).

En la Ilustración 4-9, la ventana reconoce una frase que se encuentra en el árbol de conocimiento que denota una clasificación "*Natural Language Processing*", y actualiza sus valores realizando una votación.

ClassOnto: Funcionamiento

• Texto:

Se utilizará técnicas de procesamiento de lenguaje natural, para la clasificación de textos y reconocimiento de patrones de interés. De igual manera se hará uso de minería de datos para la obtención de corpus.

Si representa una clasificación

Natural Language Processing: 0 -> Natural Language Processing: 1

Ilustración 4-9. Funcionamiento del algoritmo (9).

Se realiza nuevamente el recorrido de la ventana en la Ilustración 4-10, hasta que aparezca una nueva frase a reconocer.

ClassOnto: Funcionamiento

• Texto:

Se utilizará técnicas de procesamiento de lenguaje natural, para la clasificación de textos y reconocimiento de patrones de interés. De igual manera se hará uso de minería de datos para la obtención de corpus.

No representa una clasificación

Ilustración 4-10. Funcionamiento del algoritmo (10).

En Ilustración 4-11, la ventana reconoce una frase que se encuentra en el árbol de conocimiento y que denota una clasificación "*Text Classification*"; procediendo a actualizar sus valores realizando una votación. Esta clasificación a su vez pertenece a una clasificación más general "*Natural Language Processing*" por lo que se realiza una propagación del voto, de acuerdo a un factor, y dado que la clasificación general cuenta con voto previamente realizado, se le suma el valor de la propagación.

ClassOnto: Funcionamiento

• Texto:

Se utilizará técnicas de procesamiento de lenguaje natural, para la clasificación de textos y reconocimiento de patrones de interés. De igual manera se hará uso de minería de datos para la obtención de corpus.

• *Natural Language Procesing: 1*
  • *Text Classification: 0*

->

• *Natural Language Procesing: 1.5*
  • *Text Classification: 1*

Ilustración 4-11. Funcionamiento del algoritmo (11).

En la Ilustración 4-12, se muestra otra frase reconocida que denota una clasificación "*Pattern Recognition*", que a su es parte de las clasificaciones generales "*Natural Language Processing*" e "*Imaging Processing*", por lo que para ambas clasificaciones generales se vota con su respectivo factor de propagación. Y dado que "*Natural Language Processing*" ya tiene votos, se suman los nuevos realizados.

ClassOnto: Funcionamiento

• Texto:

Se utilizará técnicas de procesamiento de lenguaje natural, para la clasificación de textos y reconocimiento de patrones de interés. De igual manera se hará uso de minería de datos para la obtención de corpus.

- *Natural Language Processing: 1.5*
  - *Pattern Recognition: 0*
- *Imaging Processing: 0*
  - *Pattern Recognition: 0*

->

- *Natural Language Processing: 2.0*
  - *Pattern Recognition: 1*
- *Imaging Processing: 0.5*
  - *Pattern Recognition: 1*

Ilustración 4-12. Funcionamiento del algoritmo (12).

En la Ilustración 4-13, continúa con la búsqueda recorriendo la ventana, hasta llegar al final del documento.

ClassOnto: Funcionamiento

• Texto:

Se utilizará técnicas de procesamiento de lenguaje natural, para la clasificación de textos y reconocimiento de patrones de interés. De igual manera se hará uso de minería de datos para la obtención de corpus.

No representa una clasificación

Ilustración 4-13. Funcionamiento del algoritmo (13).

En la Ilustración 4-14, sucede el caso sobre algún término que no estará en el árbol de conocimientos, sin embargo, por su frecuencia de aparición en el documento, puede indicar que denota algún significado.

ClassOnto: Funcionamiento

• Texto:

Se utilizará técnicas de procesamiento de lenguaje natural, para la clasificación de textos y reconocimiento de patrones de interés. De igual manera se hará uso de minería de datos para la obtención de corpus.

No representa una clasificación, PERO se asigna al diccionario de palabras por definir (posiblemente tenga una frecuencia muy alta de aparición)

Ilustración 4-14. Funcionamiento del algoritmo (14).

Hasta la Ilustración 4-14 se ha terminado con el recorrido de la ventana, por lo que ya no queda más texto que analizar.

4.3. REQUERIMIENTOS

El establecimiento de requerimientos se divide en dos partes:

- Funcionales: describen en lenguaje natural, la funcionalidad de alto nivel del sistema.
- No funcionales: describe los requerimientos en el nivel de usuario que no están relacionado en forma directa con la funcionalidad. Esto incluye el desempeño, la seguridad, la modificabilidad, el manejo de errores, la plataforma y el ambiente físico.

4.3.1. *REQUERIMIENTOS FUNCIONALES*

- CLASSONTO debe clasificar la publicación especializada bajo la ontología ACM para las ciencias de la computación.
- CLASSONTO podrá aceptar un formato de archivo TXT que describa la ontología, con el cual las publicaciones se quieren clasificar.
- CLASSONTO podrá cambiar el factor de propagación de la votación.
- CLASSONTO deberá mostrar los valores de votación para cada término de la ontología por publicación analizada, y guardarlos en formato CSV.
- CLASSONTO deberá mostrar los términos no conocidos por cada publicación analizada.

4.3.2. *REQUERIMIENTOS NO FUNCIONALES*

- CLASSONTO debe permitir la fácil adaptación y adecuación de nuevos árboles de conocimientos.
- CLASSONTO se programará en Python.

4.3.3. *DIAGRAMA DE ACTIVIDADES DEL SISTEMA*

La Ilustración 4-15 se muestra el diagrama de actividad que se usa para representar el flujo de datos de un sistema.

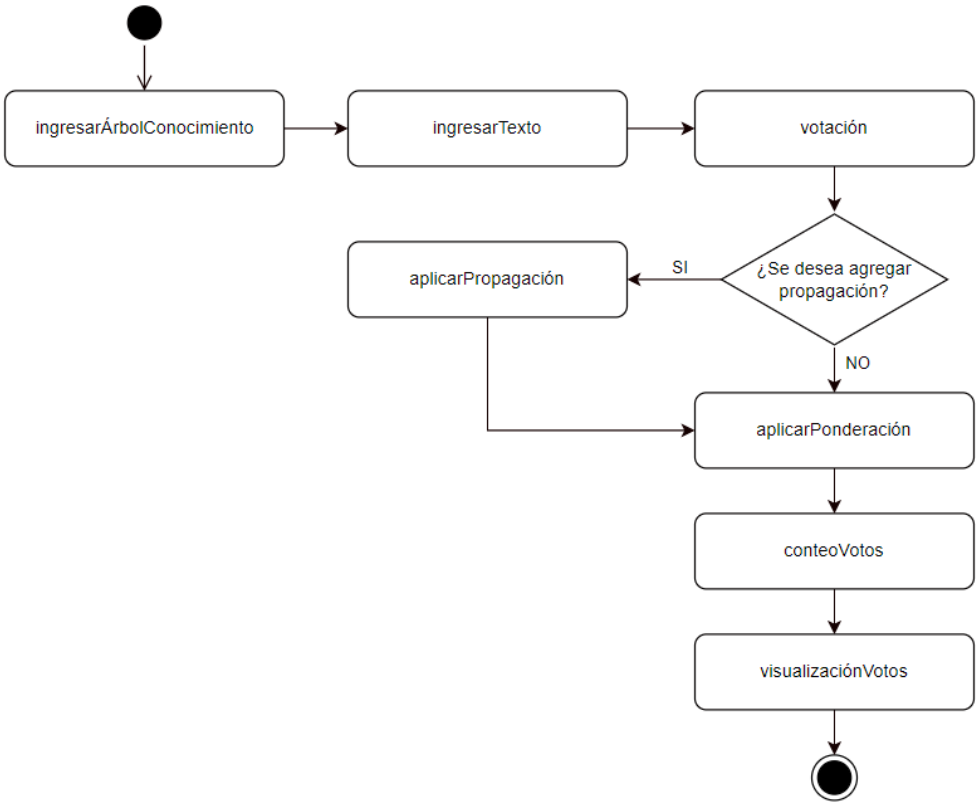


Ilustración 4-15. Diagrama de actividades del sistema.

4.4. REFERENCIAS DEL CAPÍTULO

[1] R. S. Pressman, *Ingeniería del software. Un enfoque práctico*, Séptima ed. McGraw Hill, 2010.

[2] J. A. Senn, *Análisis y diseño de sistemas de información*, Segunda ed. McGraw Hill, 1992.

[3] L. Michán and I. Muñoz-Velasco, "Cienciometría para ciencias médicas: definiciones, aplicaciones y perspectivas," *Investig. en Educ. Médica*, vol. 2, no. 6, pp. 100–106, Apr. 2013, doi: 10.1016/s2007-5057(13)72694-2.

5. Desarrollo y Resultados

Acrónimos del capítulo

LPC	Linear Predictive Coding (Codificación predictiva lineal)
PDS	Procesador Digital de Señales
DSP	Digital Signal Processor (Procesador digital de señales)
MEMS	Microelectromechanical systems (Sistemas microelectromecánicos)
HCI	Human Centered Interface (Interfaz centrada al humano)
CIC	Centro de Investigación en Computación
ACM	Association for Computing Machinery (Asociación de Maquinaria Computacional)
TP	True Positive (Verdaderos Positivos)
TN	True Negative (Verdaderos Negativos)
FP	False Positive (Falsos Positivos)
FN	False Negative (Falsos Negativos)
F1	Una medida que combina precisión y recuperación es la media armónica de precisión y recuperación.

5.1. INTRODUCCIÓN

En este capítulo se describe el desarrollo y los resultados que se tuvieron durante la realización de este trabajo. Se muestra el desarrollo de lo planteado en el capítulo de análisis y diseño, junto con los resultados en tablas y gráficas.

5.1.1. *OBTENCIÓN DE LAS CLASIFICACIONES*

Antes de realizarse la poda, CLASSONTO encontraba muchas clasificaciones, con un promedio superior a la 120 por tesis, evidentemente con clasificaciones en la que su votación era insignificante (valores muy bajos). La elección de quedarse con 10 clasificaciones va acorde a la facilidad de describir una tesis con 10 clasificaciones, siendo que en muchos casos para la descripción de un artículo de investigación se recomiendan entre 3 y 7 palabras clave (keywords).

5.1.2. *TABLA DE TERMINOLOGÍA POR RECONOCER*

Como lo comentado en la sección "4.2.3. Poda de términos no reconocidos en el corpus", las palabras que no realizaron votación a las clasificaciones se agrupan en una sola tabla, al igual que en cuál y cuantas tesis están presentes. Esto permite identificar vocabulario compartido, que bien puede representar términos en el consenso del ámbito de la computación.

word	value	files	count_files
conductance	93	{'FigueroaAguilarIsaac_2019', 'HerreraHernandezDiego_2014', 'ProaCoronadoSergio_2015', 'LakeMocte	6
testor	92	{'VazquezTorresFernando_2008', 'GarciaMartinezEdgarAlfonso_2014', 'GonzalezGuevaraVictorIvan_20	6
glioma	91	{'CastroLopezRoberto_2013', 'HerreraMagañaJorgeAlberto_2016', 'VelazquezRodriguezJoseLuis_2018',	6
liwc	90	{'HernandezCastañedaAngel_2017', 'JuarezGambinoJoelOmar_2019', 'RiveraCamachoRamon_2017', 'A	6
irm	90	{'CaroVasquezJoseRoberto_2017', 'MartinezFelipeMiguelDeJesus_2018', 'GonzalezBonillaJavier_2017'	6
hipocampo	88	{'ValleChavezAbel_2012', 'GutierrezDeLaPazOmarAlfonso_2017', 'SantiagoNievesRodrigoRuben_2015'	6
quincena	88	{'GuerreroHernandezMauricioIvan_2015', 'CastilloOrtaMiguelAngel_2015', 'MartinezHernandezVictor	6
uint	88	{'MironBernalMiguelAngel_2008', 'RiveraZamarripaLuisAlberto_2019', 'LunaNuñezBrayan_2014', 'Nava	6
apoderar	87	{'PosadasDuranJuanPabloFrancisco_2011', 'ChiChimManuelAntonio_2008', 'SandovalFloresFranciscoG	6
pyramidal	82	{'ValadezGodinezSergio_2018', 'ArgüellesCruzAmadeoJose_2007', 'FigueroaAguilarIsaac_2019', 'Alfaro	6
mpc	81	{'AriasOlivasMarvinRene_2008', 'RangelGonzalezJosue_2009', 'EstradaHernandezJoseAbraham_2016',	6
asertividad	81	{'MarquezOliveraMoisesVicente_2012', 'BarceloAlonsoGrettel_2010', 'RiveraAzamarJuanCarlos_2012',	6
biomecánicos	79	{'JimenezGarciaGregorioArturo_2016', 'ConchaGomezPaulaDenisse_2018', 'GarzaRodriguezAlejandro_	6
dmt	78	{'GuevaraLopezPedro_2004', 'AntonioMendezRaul_2011', 'LagunaSanchezGerardoAbel_2010', 'LopezM	6
volp	76	{'CansecoRodriguezMariaTeresa_2012', 'MartinezOrtuñoJoel_2014', 'BejaranoLeyvaCruzAlonso_2008',	6
specie	75	{'ViguerasVelazquezMidoryEsmeralda_2016', 'AlvaradoGutierrezJesusAlexander_2016', 'CelisPorrasJe	6
jasa	75	{'CabreraAlvarezErickNicolas_2012', 'RomeroAlvarioAdolfo_2002', 'MontañoSanchezCesarEdgar_2016',	6
ventricular	73	{'VillegasLopezCristianEduardo_2014', 'AlanisTamezMarianaDayanara_2018', 'TamboneroXixitlaVianne	6
xps	73	{'RamirezLazoCristobal_2016', 'GomezCondeAlejandro_2013', 'SanchezPerezLuisAlejandro_2011', 'Alor	6
coroides	72	{'RangelFajardoVictorManuel_2018', 'VillalobosCastaldiFabiolaMiroslaba_2011', 'AlbortanteMoratoCe	6
schreckenberg	71	{'AlvarezMendezJonatan_2014', 'LunaBenosoBenjamin_2006', 'SantanaRivasIsmaelDeJesus_2009', 'Silc	6
rpg	70	{'AlonsoLaverniaMariaDeLosAngeles_2006', 'VargasDeBasterraRicardo_2006', 'HortaMendozaJuanMan	6
mindstorm	70	{'MoralesVazquezBrendaVeronica_2014', 'BuchanCastilloEnrique_2011', 'ToncheGarciaRonnyJose_201	6
boxplot	70	{'VilchisMompalaRodolfoAntonio_2013', 'GutierrezCeballosManuel_2017', 'UriarteArciaAbrilValeria_2	6
rtd	69	{'CarreraTrejoJorgeVictor_2010', 'MontañoSanchezCesarEdgar_2016', 'AlfaroPonceMariel_2015', 'Parec	6
lovaina	69	{'LandassuriMorenoVictorManuel_2006', 'BautistaThompsonErnestoFrancisco_2005', 'RamirezAmaroK	6
axonal	68	{'BarronFernandezRicardo_2006', 'ValadezGodinezSergio_2018', 'GuevaraMartinezElizabeth_2016', 'Alf	6
dominating	68	{'GarciaDiazJesus_2013', 'NavarroZayasRodolfo_2014', 'GarciaDiazJesus_2017', 'CruzTrejoAriana_2012',	6
rssi	66	{'RiveraAzamarJuanCarlos_2012', 'GarciaGonzalezEmmanuel_2018', 'PinedaBriseñoAnabel_2013', 'Avil	6
cósmico	66	{'CallejasRamosAlejandroIvan_2016', 'PriegoPerezFaustoPavel_2012', 'HerreraLopezMariaGuadalupe_	6
psf	66	{'LopezEnriquezJuanCarlos_2014', 'CallejasRamosAlejandroIvan_2016', 'GarciaOrdazJoseRaul_2010', 'C	6
homicidio	66	{'GutierrezCeballosManuel_2017', 'CastilloOrtaMiguelAngel_2015', 'JuarezGambinoJoelOmar_2008', 'R	6
reprint	66	{'VazquezTorresFernando_2008', 'ClempnerKerikJulioBernardo_2006', 'HernandezMartinezErick_2017'	6

Ilustración 5-1. Fragmento de tabla de términos no reconocidos.

En la Ilustración 5-1se puede observar distintas palabras que posiblemente puedan votar por una clasificación como lo son: "mpc", "hipocampo", "uint", "testor", "conductance", etc. De esta forma se han agregado los términos: "memorias asociativas", "LPC", "PDS", "DSP", "MEMS", etc., al árbol de conocimientos.

5.1.2.1. Inferencia de clasificaciones en terminología no reconocida

La inferencia de clasificaciones permite sugerir terminología relacionada como un proceso semiautomático.

word	valu	files	count_files	inferenceWords_HighMedium	inferenceAncestorWords_HighMedium
paráfrasis	242	{'GomezAdornoHelenaMontserrat_2017', 'He	12	['11.3.1. natural language processing']	['11.3.1. natural language processing', '11.3. artificial intelligence']
senseval	172	{'RiveraLozaGabriela_2003', 'ViverosJimenez	11	['11.3.1. natural language processing']	['11.3.1. natural language processing', '11.3. artificial intelligence']
semcor	141	{'LedoMezquitaYoel_2006', 'ViverosJimenezI	10	['11.3.1. natural language processing']	['11.3.1. natural language processing', '11.3. artificial intelligence']
correferencia	259	{'AlonsoCastroJoseAdriel_2017', 'LibradoJacc	9	['11.3.1. natural language processing']	['11.3.1. natural language processing', '11.3. artificial intelligence']
paraphrase	191	{'LedenevaYuliaNikolaevna_2008', 'RiosGaor	8	['11.3.1. natural language processing']	['11.3.1. natural language processing', '11.3. artificial intelligence']
emotiv	168	{'FigueroaAguilarIsaac_2019', 'GuadarramaRe	7	['8.5.5.8. clustering and classification', '12.4.1.7. imaging']	['8.5.5.8. clustering and classification', '8.5.5. retrieval tasks and goals', '12.4.1. computational biology', '12.4.1.7. imaging']
anafórico	264	{'GomezAdornoHelenaMontserrat_2017', 'Ol	7	['11.3.1. natural language processing']	['11.3.1. natural language processing', '11.3. artificial intelligence']
cifrador	91	{'MartinezHernandezJuanManuel_2014', 'Rui	7	['9.1.2. symmetric cryptography and hash functions']	['9.1.2. symmetric cryptography and hash functions', '9.1. cryptography']
updrs	102	{'RodriguezJordanGabrielDeJesus_2016', 'Co	6	['12.4. life and medical sciences']	['12.4. life and medical sciences']
pronación	120	{'AcostaArenasAnaRosa_2018', 'ConchaGome	6	['12.4. life and medical sciences']	['12.4. life and medical sciences']
spsoo	136	{'OlivasZazuetaOmarAlejandro_2006', 'Juare:	6	['11.3.1. natural language processing']	['11.3.1. natural language processing', '11.3. artificial intelligence']
biomecánicos	79	{'RodriguezJordanGabrielDeJesus_2016', 'Co	6	['12.4. life and medical sciences']	['12.4. life and medical sciences']
desambiguador	61	{'TorresRamosSulema_2010', 'DiazRangelIsm	6	['11.3.1. natural language processing']	['11.3.1. natural language processing', '11.3. artificial intelligence']
jts	61	{'SanchezPerezLuisAlejandro_2011', 'Carrera	6	['8.3.3.2. geographic information systems']	['8.3.3. spatial-temporal systems', '8.3.3.2. geographic information systems']
daofso	71	{'EscamillaBouchanImelda_2016', 'ZhilaAlisa_	5	['11.3.1. natural language processing']	['11.3.1. natural language processing', '11.3. artificial intelligence']
diomso	53	{'EscamillaBouchanImelda_2016', 'JuarezGar	5	['11.3.1. natural language processing']	['11.3.1. natural language processing', '11.3. artificial intelligence']
microarreglos	68	{'TorresCalderonAndrea_2016', 'RomanGodir	5	['11.4.1.1. supervised learning']	['11.4.1.1. supervised learning', '11.4.1. learning paradigms']
ngram	83	{'LedenevaYuliaNikolaevna_2008', 'SanchezP	5	['11.3.1. natural language processing']	['11.3.1. natural language processing', '11.3. artificial intelligence']
supinación	115	{'ConchaGomezPaulaDenisse_2018', 'Doming	5	['12.4. life and medical sciences']	['12.4. life and medical sciences']
spatio	71	{'FigueroaAguilarIsaac_2019', 'LakeMoctezun	5	['8.5.5.8. clustering and classification']	['8.5.5.8. clustering and classification', '8.5.5. retrieval tasks and goals']
fweef	51	{'RamosMarquezJuanCarlos_2017', 'LibradoJa	5	['8.3.2.5. social networking sites', '10.3.1.5. social networks', '10.3.3.5. social networking sites', '8.4.4.2. social networks']	['10.3.1. collaborative and social computing theory, concepts and paradigms', '10.3.3. collaborative and social computing systems and tools', '8.3.2.5. social networking sites', '10.3.1.5. social networks', '8.3.2. collaborative and social computing systems and tools', '8.4.4.2. social networks', '10.3.3.5. social networking sites', '8.4.4. web applications']
geosparql	57	{'CabreraRiveraLuis_2014', 'ZarateEscobedoR	5	['8.4.7.1. semantic web description languages', '8.5.1.6. ontologies', '12.2.8. enterprise ontologies, taxonomies and vocabularies']	['8.4.7. web data description languages', '12.2.8. enterprise ontologies, taxonomies and vocabularies', '8.4.7.1. semantic web description languages', '8.5.1.6. ontologies', '8.5.1. document representation', '12.2. enterprise computing']
masc	107	{'OlivasZazuetaOmarAlejandro_2006', 'LugoG	5	['11.3.1. natural language processing']	['11.3.1. natural language processing', '11.3. artificial intelligence']
desinencia	77	{'AguilarGaliciaHonorato_2012', 'CalvoCastro	4	['11.3.1. natural language processing']	['11.3.1. natural language processing', '11.3. artificial intelligence']

Ilustración 5-2. Sugerencia de clasificaciones en terminología no reconocida.

En la Ilustración 5-2 se observa las posibles clasificaciones a las que podrían pertenecer la terminología no relacionada, por ejemplo: el término "paráfrasis" a la clasificación de "11.3.1. Natural Language Processing" de igual manera con el término "anafórico". A su vez el término "emotiv" sugiere las clasificaciones "8.5.5.8. Clustering and Classification" y "12.4.1.7. Imaging".

Debido a que estas sugerencias son a partir de tesis clasificadas por CLASSONTO similarmente, estas sugerencias se pueden filtrar, teniendo la posibilidad de tomar las 10 clasificaciones obtenidas por el sistema, o acotándolas por su etiqueta de importancia (Alta, Media, Baja). En la Ilustración 5-2 se acotó a clasificaciones de importancia alta y media, no tomando en cuenta las de baja importancia.

Las clasificaciones al estar en una estructura jerárquica, sus ancestros posiblemente también brinden información importante, por lo que también se puede tener mayor

información al añadir el ancestro inmediatamente superior a la clasificación sugerida, aunque teniendo como limitante no llegar al nivel de clasificación más superior, para no generalizar el concepto del término; como sucede con el término "pronación" y su clasificación "12.4. Life and Medical Sciences", siendo el ancestro inmediato superior la clasificación "12. Applied Computing".

5.1.3. *CLASIFICACIÓN Y VALIDACIÓN*

Se mostrarán algunas tesis con sus clasificaciones predichas. También se mostrará las métricas de validación con los 3 modos propuestos (bondadoso, estricto y onehot) y con las métricas clásicas *F1, precision, recall.*

5.1.3.1. CLASIFICACIONES DE TESIS

A continuación, se muestran algunas tesis con su respectivo título y resumen, así como la tabla de clasificación que se obtuvo, y su comparación con el *Golden Standard.*

El código de colores representa lo siguiente:

- Valores resaltados en rojo significan que el algoritmo no encontró la clasificación.
- Valores resaltados en verde significa que el algoritmo encontró la clasificación y que la etiqueta "importancia" fue la correcta.
- Valores resaltados en amarillo significa que el algoritmo encontró la clasificación y que la etiqueta "importancia" no coincide con el *Golden Standard*.

La columna "y_true" representa el *Golden Standard*, mientras que la columna "y_pred" representa la columna de valores predichos por el algoritmo.

La columna "value" es la votación que recibió la clasificación, y "class_level" el nivel de la clasificación.

La columna "classification" es la viñeta u orden que pertenece la clasificación, mientras que la columna "Title" es el nombre de la clasificación como aparece en la ontología ACM (en idioma inglés).

5.1.3.1.1. *HERNÁNDEZ HERNÁNDEZ GERARDO 2019*

Título: Hybrid neural networks with morphological neurons and perceptrons.

ABSTRACT

In (*Machine Learning (ML)*), there are several architectures of neural networks. Architectures that for decades have been using the same computing units for pattern classification. Within these units, we find the Perceptron and the morphological neurons. Due to their theoretical sustenance and practical results, the Perceptron is the most popular element in neural network architectures, such as Multilayer Perceptron (MLP), Deep Neural Network (DNN), Convolutional Neural Network (CNN), among others.

However, it has been proved that neural network training whom includes Perceptron have several disadvantages, such as a high number of training parameters, slow training algorithms Wilson and Martinez [56], and the generated hyper-planes are not the best Glorot and Bengio [14], Bouzerdoum [8].

In this work, we focus on researching new and different computation units which allow us to reduce the complexity of neural network models, reduce training time and improve the generated hyperplanes.

With the aforementioned, two new hybrid neural network architectures are proposed, which combine perceptrons and morphological neurons. The first architecture is the Linear-Morphological Neural Network (LMNN), and the second is Morphological-Linear Neural Network (MLNN). Both architectures are trained by Stochastic Gradient Descent (SGD).

The first result shows that a single layer of morphological neurons has a greater capacity to segment 2D space in different regions than a layer of perceptrons. The second result shows that it is possible to use morphological neurons as feature extractors and, on average, this architecture requires a smaller number of training parameters than its counterparts, and at the same time, it obtains better classification percentages.

Ilustración 5-3. Captura de resumen de tesis (1).

Y las clasificaciones que reconoce el sistema.

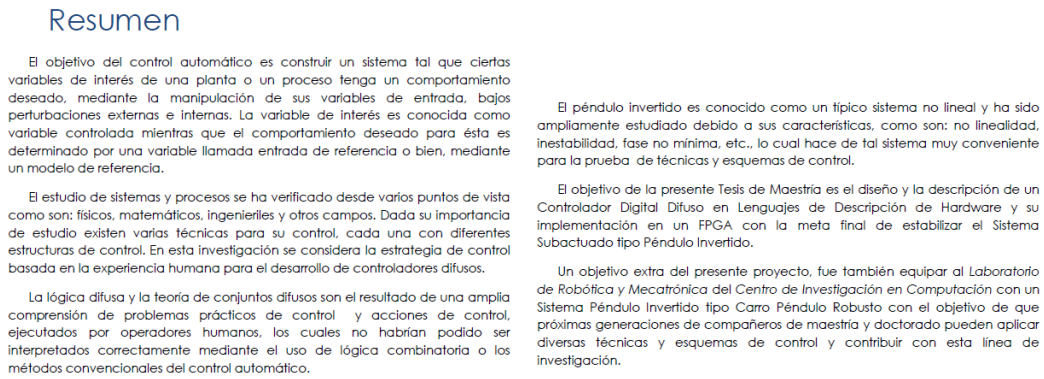
Tabla 5-1. Tabla de clasificación de tesis (1).

Class	*Title*	*Value*	*Class level*	*Y pred*	*Y true*
11.4.3.3.	neural networks	10.9440	4	High	High
11.4.1.1.	supervised learning	6.3917	4	Medium	High
11.4.3.6.1.	perceptron algorithm	4.5802	5	Medium	Low
8.5.5.8.	clustering and classification	3.1536	4	Low	
11.4.3.	machine learning approaches	1.9934	3	Low	
3.1.3.	distributed architectures	1.8583	3	Low	
11.4.1.	learning paradigms	1.1632	3	Low	
11.4.	machine learning	1.0063	2	Low	
11.4.3.2.1.	support vector machines	0.7883	5	Low	
11.4.3.6.	learning linear models	0.7633	4	Low	
3.1.4.1.	neural networks	0	4		Low

Se puede observar en Tabla 5-1, que con las clasificaciones agregadas por el algoritmo (clasificaciones sin color) se puede inferir que la tesis habla acerca de más aproximaciones del aprendizaje de máquina, y no solo de redes neuronales y perceptrón, como lo indica el título de la tesis.

5.1.3.1.2. *FLORES CORTES ANDRÉS 2013*

Título: Control inteligente de un péndulo invertido y su implementación sobre FPGA



Resumen

El objetivo del control automático es construir un sistema tal que ciertas variables de interés de una planta o un proceso tenga un comportamiento deseado, mediante la manipulación de sus variables de entrada, bajos perturbaciones externas e internas. La variable de interés es conocida como variable controlada mientras que el comportamiento deseado para ésta es determinado por una variable llamada entrada de referencia o bien, mediante un modelo de referencia.

El estudio de sistemas y procesos se ha verificado desde varios puntos de vista como son: físicos, matemáticos, ingenieriles y otros campos. Dada su importancia de estudio existen varias técnicas para su control, cada una con diferentes estructuras de control. En esta investigación se considera la estrategia de control basada en la experiencia humana para el desarrollo de controladores difusos.

La lógica difusa y la teoría de conjuntos difusos son el resultado de una amplia comprensión de problemas prácticos de control y acciones de control, ejecutados por operadores humanos, los cuales no habrían podido ser interpretados correctamente mediante el uso de lógica combinatoria o los métodos convencionales del control automático.

El péndulo invertido es conocido como un típico sistema no lineal y ha sido ampliamente estudiado debido a sus características, como son: no linealidad, inestabilidad, fase no mínima, etc., lo cual hace de tal sistema muy conveniente para la prueba de técnicas y esquemas de control.

El objetivo de la presente Tesis de Maestría es el diseño y la descripción de un Controlador Digital Difuso en Lenguajes de Descripción de Hardware y su implementación en un FPGA con la meta final de estabilizar el Sistema Subactuado tipo Péndulo Invertido.

Un objetivo extra del presente proyecto, fue también equipar al *Laboratorio de Robótica y Mecatrónica* del *Centro de Investigación en Computación* con un Sistema Péndulo Invertido tipo Carro Péndulo Robusto con el objetivo de que próximas generaciones de compañeros de maestría y doctorado pueden aplicar diversas técnicas y esquemas de control y contribuir con esta línea de investigación.

Ilustración 5-4. Captura de resumen de tesis (2).

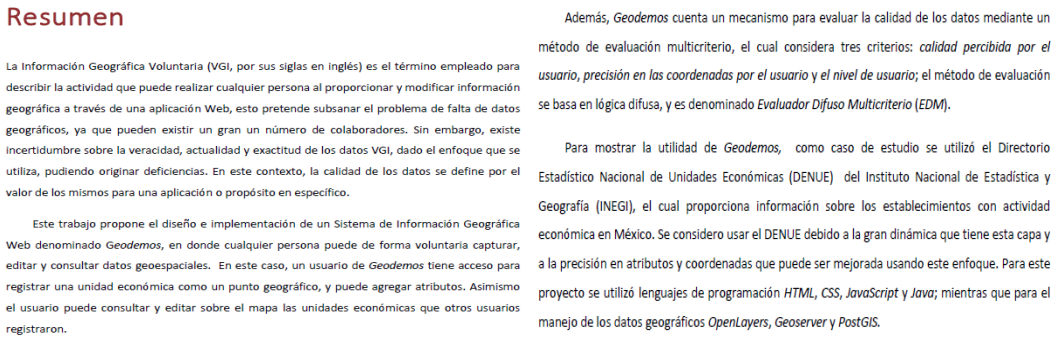
Y las clasificaciones que reconoce el sistema.

Tabla 5-2. Tabla de clasificación de tesis (2).

Class	*Title*	*Value*	*Class level*	*Y pred*	*Y true*
3.2.2.2.	robotic control	7.857	4	High	Low
2.3.6.	reconfigurable logic and fpgas	6.455	3	High	High
2.3.6.3.	programmable logic elements	6.385	4	High	High
11.3.2.5.	vagueness and fuzzy logic	5.59	4	High	Medium
3.1.4.2.	reconfigurable computing	5.229	4	Medium	
3.2.2.	robotics	2.741	3	Medium	
2.6.2.	hardware description languages and compilation	2.443	3	Low	High
2.3.	integrated circuits	1.491	2	Low	
7.3.	mathematical software	1.134	2	Low	
5.2.1.	general programming languages	1.076	3	Low	
11.3.5.2.	computational control theory	0	4		Low

5.1.3.1.3. *VAZQUEZ OROPEZA JONATHAN 2016*

Título: Información geográfica voluntaria para la actualización y validación de conjuntos de datos geoespaciales.



Resumen

La Información Geográfica Voluntaria (VGI, por sus siglas en inglés) es el término empleado para describir la actividad que puede realizar cualquier persona al proporcionar y modificar información geográfica a través de una aplicación Web, esto pretende subsanar el problema de falta de datos geográficos, ya que pueden existir un gran un número de colaboradores. Sin embargo, existe incertidumbre sobre la veracidad, actualidad y exactitud de los datos VGI, dado el enfoque que se utiliza, pudiendo originar deficiencias. En este contexto, la calidad de los datos se define por el valor de los mismos para una aplicación o propósito en específico.

Este trabajo propone el diseño e implementación de un Sistema de Información Geográfica Web denominado *Geodemos*, en donde cualquier persona puede de forma voluntaria capturar, editar y consultar datos geoespaciales. En este caso, un usuario de *Geodemos* tiene acceso para registrar una unidad económica como un punto geográfico, y puede agregar atributos. Asimismo el usuario puede consultar y editar sobre el mapa las unidades económicas que otros usuarios registraron.

Además, *Geodemos* cuenta un mecanismo para evaluar la calidad de los datos mediante un método de evaluación multicriterio, el cual considera tres criterios: *calidad percibida por el usuario*, *precisión en las coordenadas por el usuario* y *el nivel de usuario*; el método de evaluación se basa en lógica difusa, y es denominado *Evaluador Difuso Multicriterio (EDM)*.

Para mostrar la utilidad de *Geodemos*, como caso de estudio se utilizó el Directorio Estadístico Nacional de Unidades Económicas (DENUE) del Instituto Nacional de Estadística y Geografía (INEGI), el cual proporciona información sobre los establecimientos con actividad económica en México. Se considero usar el DENUE debido a la gran dinámica que tiene esta capa y a la precisión en atributos y coordenadas que puede ser mejorada usando este enfoque. Para este proyecto se utilizó lenguajes de programación *HTML*, *CSS*, *JavaScript* y *Java*; mientras que para el manejo de los datos geográficos *OpenLayers*, *Geoserver* y *PostGIS*.

Ilustración 5-5. Captura de resumen de tesis (3).

Y las clasificaciones que reconoce el sistema.

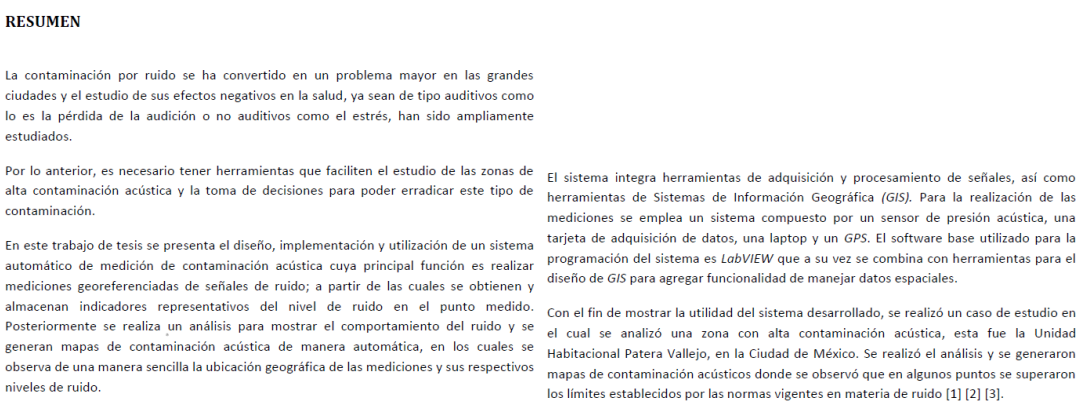
Tabla 5-3. Tabla de clasificación de tesis (3).

Class	*Title*	*Value*	*Class level*	*Y pred*	*Y true*
8.3.3.2.	geographic information systems	13.707	4	High	High
8.3.3.	spatial-temporal systems	2.7800	3	Low	
8.1.3.2.	database query processing	1.4613	4	Low	
11.3.2.5.	vagueness and fuzzy logic	1.4109	4	Low	
10.1.1.	hci design and evaluation methods	1.4109	3	Low	
10.5.6.	visualization design and evaluation methods	1.4109	3	Low	Low
10.6.3.	accessibility design and evaluation methods	1.4109	3	Low	
5.2.1.1.11.	multiparadigm languages	1.3353	5	Low	
8.3.3.1.	location based services	0.9574	4	Low	Medium
8.4.4.	web applications	0.8804	3	Low	
10.2.1.1.	user interface design	0	4		Low
10.5.2.3.	geographic visualization	0	4		High

Aun cuando no haya identificado todas las clasificaciones del *Golden Standard* en Tabla 5-3, se puede inferir que por las clasificaciones "10.1.1. hci design and evaluation methods", "10.5.6 visualization design and evaluation methods" y "10.6.3. accessibility design and evaluation methods" se desarrolló una interfaz de usuario, con su visualización y sus pruebas de accesibilidad.

5.1.3.1.4. *QUINTERO PEREZ GUILLERMO 2013*

Título: Sistema móvil georreferenciado para la medición y análisis de ruido ambiental.



RESUMEN

La contaminación por ruido se ha convertido en un problema mayor en las grandes ciudades y el estudio de sus efectos negativos en la salud, ya sean de tipo auditivos como lo es la pérdida de la audición o no auditivos como el estrés, han sido ampliamente estudiados.

Por lo anterior, es necesario tener herramientas que faciliten el estudio de las zonas de alta contaminación acústica y la toma de decisiones para poder erradicar este tipo de contaminación.

En este trabajo de tesis se presenta el diseño, implementación y utilización de un sistema automático de medición de contaminación acústica cuya principal función es realizar mediciones georeferenciadas de señales de ruido; a partir de las cuales se obtienen y almacenan indicadores representativos del nivel de ruido en el punto medido. Posteriormente se realiza un análisis para mostrar el comportamiento del ruido y se generan mapas de contaminación acústica de manera automática, en los cuales se observa de una manera sencilla la ubicación geográfica de las mediciones y sus respectivos niveles de ruido.

El sistema integra herramientas de adquisición y procesamiento de señales, así como herramientas de Sistemas de Información Geográfica *(GIS)*. Para la realización de las mediciones se emplea un sistema compuesto por un sensor de presión acústica, una tarjeta de adquisición de datos, una laptop y un *GPS*. El software base utilizado para la programación del sistema es *LabVIEW* que a su vez se combina con herramientas para el diseño de *GIS* para agregar funcionalidad de manejar datos espaciales.

Con el fin de mostrar la utilidad del sistema desarrollado, se realizó un caso de estudio en el cual se analizó una zona con alta contaminación acústica, esta fue la Unidad Habitacional Patera Vallejo, en la Ciudad de México. Se realizó el análisis y se generaron mapas de contaminación acústicos donde se observó que en algunos puntos se superaron los límites establecidos por las normas vigentes en materia de ruido [1] [2] [3].

Ilustración 5-6. Captura de resumen de tesis (4).

Y las clasificaciones que reconoce el sistema.

Tabla 5-4. Tabla de clasificación de tesis (4).

Class	*Title*	*Value*	*Class level*	*Y pred*	*Y true*
8.3.3.2.	geographic information systems	10.1024	4	High	High
8.3.3.1.	location based services	6.1845	4	Medium	
8.3.3.5.	global positioning systems	5.2051	4	Medium	
8.3.3.	spatial-temporal systems	3.8906	3	Medium	
2.9.3.5.	signal integrity and noise analysis	5.0504	4	Medium	High

2.2.1.1.	digital signal processing	4.4834	4	Medium	High
8.1.1.1.	relational database model	2.7822	4	Low	
8.3.	Information systems applications	1.8225	2	Low	
5.2.1.1.8.	Data Flow languages	1.6334	5	Low	
8.3.2.	collaborative and social computing systems and tools	1.5622	3	Low	Medium

Con la información de la clasificación en la Tabla 5-4, podemos inferir que el autor no ocupó un DSP para la adquisición de señales de ruido, y que posiblemente no se basó en un repositorio para la adquisición de señales de ruido, sino que posiblemente el hizo su propia base de datos.

5.1.3.2. VALIDACIÓN DE CLASIFICACIONES

La Tabla 5-5 muestra las métricas de de validación para las 100 tesis que previamente se habían clasificado (Golden Standard). Teniendo en cuenta que los valores máximos que se pueden alcanzar para los modos de validación propuestos (estricto, bondadoso y onehot) son de 1, 1.5 y 1 respectivamente.

Tabla 5-5. Métricas de validación *precision recall, F1* y propuestas.

	Estricto	*Bondadoso*	*Onehot*	*Precision*	*Recall*	*F1*
Promedio	0.742	0.989	0.81	0.334	0.81	0.466
Máximo	1	1.5	1	0.7	1	0.823
Mínimo	0.264	0.357	0.286	0.1	0.29	0.153
Moda	0.9	1.166	1	0.3	1	0.461
Mediana	0.74	1	0.8	0.3	0.8	0.461

A continuación, en la Tabla 5-6 se muestran los valores de los modos de validación para cada tesis, junto con los valores de *precision*, *recall* y *F1*.

Tabla 5-6. Valores de métricas de validación por tesis.

filename	*Strict*	*Kind*	*Onehot*	*Precision*	*Recall*	*F1*
AcevedoMosquedaMariaElena_2006	*0.95*	*1.3333*	*1*	*0.3*	*1*	*0.4615*
AguilarGaliciaHonorato_2012	*0.9*	*1.1667*	*1*	*0.3*	*1*	*0.4615*

AlbortanteMoratoCecilia_2009	0.283	0.3333	0.333	0.1	0.33	0.1538
AlcazarSilvaEliezer_2013	0.963	1.375	1	0.4	1	0.5714
AlonsoCastroJoseAdriel_2017	0.925	1.25	1	0.4	1	0.5714
AlvarezMendezJonatan_2014	0.667	1	0.667	0.2	0.67	0.3077
AvilaGamboaGuillermoIII_2017	0.9	1.1667	1	0.3	1	0.4615
BarceloAlonsoGrettel_2010	0.888	1.125	1	0.4	1	0.5714
BarronFernandezRicardo_2006	0.9	1.1667	1	0.3	1	0.4615
BautistaBautistaPatricia_2009	0.9	1.1667	1	0.3	1	0.4615
CabreraRiveraLuis_2018	0.771	1	0.857	0.6	0.86	0.7059
CamachoEscotoJoseJaime_2017	0.925	1.25	1	0.4	1	0.5714
CarreraTrejoJorgeVictor_2010	0.963	1.375	1	0.4	1	0.5714
CastilloMontielErandi_2009	0.5	0.75	0.5	0.2	0.5	0.2857
CatalanSalgadoEdgarArmando_2007	0.888	1.125	1	0.4	1	0.5714
CeronFigueroaSergio_2013	0.875	1.0833	1	0.6	1	0.75
CeronFigueroaSergio_2018	0.675	0.875	0.75	0.3	0.75	0.4286
CervantesRamirezAngelOmar_2012	0.356	0.5	0.375	0.3	0.38	0.3333
CoronaBermudezErendira_2020	0.97	1.4	1	0.5	1	0.6667
CortesAntonioPrometeo_2011	0.888	1.125	1	0.4	1	0.5714
CruzSilvaJacoboEmanuel_2020	0.95	1.3333	1	0.3	1	0.4615
CuevasRasgadoAlmaDelia_2003	0.57	0.8	0.6	0.3	0.6	0.4
DiazDiazJaime_2012	0.57	0.8	0.6	0.3	0.6	0.4
DuchanoyMartinezCarlosAlberto_2012	0.68	0.8	0.8	0.4	0.8	0.5333
EstradaSanchezIvan_2010	0.675	0.875	0.75	0.3	0.75	0.4286
FarfanEstradaIsmael_2012	0.5	0.75	0.5	0.2	0.5	0.2857
FernandezCidHugoIvan_2019	0.95	1.3333	1	0.3	1	0.4615
FloresCortesAndres_2013	0.758	1	0.833	0.5	0.83	0.625
FloresRoaAlberto_2010	0.74	1	0.8	0.4	0.8	0.5333
GarciaBlanquelEricka_2012	0.264	0.3571	0.286	0.2	0.29	0.2353
GarciaBlanquelEricka_2017	0.936	1.2857	1	0.7	1	0.8235
GarciaCortesMonica_2014	0.74	1	0.8	0.4	0.8	0.5333
GarciaLavanderosNorberto_2018	0.617	0.8333	0.667	0.2	0.67	0.3077
GarroLiconBeatrizAurora_2012	0.75	1.125	0.75	0.3	0.75	0.4286
GermanSotoErnesto_2002	0.97	1.4	1	0.5	1	0.6667
GodinezFernandezEduardo_2009	0.925	1.25	1	0.4	1	0.5714
GomezLuisAlexiaItzel_2019	0.638	0.75	0.75	0.3	0.75	0.4286
GonzalezGarciaAlainCesar_2007	0.45	0.5833	0.5	0.3	0.5	0.375
GutierrezSanchezAngelIvan_2015	0.37	0.5	0.4	0.2	0.4	0.2667
HernandezAcostaCindyGabriela_2018	1	1.5	1	0.3	1	0.4615
HernandezCruzEstelaYadira_2020	0.88	1.1	1	0.5	1	0.6667
HernandezHernandezGerardo_2014	0.95	1.3333	1	0.3	1	0.4615
HernandezHernandezGerardo_2019	0.675	0.875	0.75	0.3	0.75	0.4286
HerreraLopezMariaGuadalupe_2018	0.425	0.5	0.5	0.2	0.5	0.2857
HuertaMorenoGabrielOmar_2009	0.638	0.75	0.75	0.3	0.75	0.4286
IbarraRomeroMartin_2014	0.888	1.125	1	0.4	1	0.5714
IbarraVargasJoseJonathan_2009	0.925	1.25	1	0.4	1	0.5714

JimenezArandaItzael_2018	*0.888*	*1.125*	*1*	*0.4*	*1*	*0.5714*
JimenezGarciaGregorioArturo_2016	*0.9*	*1.1667*	*1*	*0.3*	*1*	*0.4615*
JuarezHipolitoJuanHumberto_2016	*0.675*	*0.875*	*0.75*	*0.3*	*0.75*	*0.4286*
JuarezMurilloCristianRemington_2012	*0.71*	*0.9*	*0.8*	*0.4*	*0.8*	*0.5333*
KolesnikovaOlga_2011	*0.71*	*0.9*	*0.8*	*0.4*	*0.8*	*0.5333*
LakeMoctezumaFranzLudwig_2019	*0.54*	*0.7*	*0.6*	*0.3*	*0.6*	*0.4*
LavinVillaMoisesEduardo_2010	*0.425*	*0.5*	*0.5*	*0.3*	*0.5*	*0.375*
LopezCardenasRodrigo_2008	*0.94*	*1.3*	*1*	*0.5*	*1*	*0.6667*
LopezPachecoMariaGuadalupe_2012	*0.888*	*1.125*	*1*	*0.4*	*1*	*0.5714*
LopezVerasteguiGermanOswaldo_2014	*0.9*	*1.1667*	*1*	*0.3*	*1*	*0.4615*
LopezYaÃ±ezItzama_2011	*0.57*	*0.8*	*0.6*	*0.3*	*0.6*	*0.4*
MarquezMolinaMiguel_2013	*0.925*	*1.25*	*1*	*0.4*	*1*	*0.5714*
MartinezCastilloJesusElohim_2012	*0.9*	*1.1667*	*1*	*0.3*	*1*	*0.4615*
MartinezCazaresEduardo_2009	*0.463*	*0.625*	*0.5*	*0.2*	*0.5*	*0.2857*
MartinezHernandezVictorManuel_2009	*0.963*	*1.375*	*1*	*0.4*	*1*	*0.5714*
MartinezNavarroJoseAngel_2018	*0.95*	*1.3333*	*1*	*0.3*	*1*	*0.4615*
MartinezValleCarlosLeon_2006	*0.5*	*0.75*	*0.5*	*0.2*	*0.5*	*0.2857*
MataRiveraMiguelFelix_2009	*0.888*	*1.125*	*1*	*0.4*	*1*	*0.5714*
MercadoCapistranPatricio_2015	*0.54*	*0.7*	*0.6*	*0.3*	*0.6*	*0.4*
MontaÃ±oSanchezCesarEdgar_2016	*0.9*	*1.1667*	*1*	*0.3*	*1*	*0.4615*
MoralesFernandezAlecxis_2009	*0.71*	*0.9*	*0.8*	*0.4*	*0.8*	*0.5333*
NovoaCataÃ±oJavier_2007	*0.71*	*0.9*	*0.8*	*0.4*	*0.8*	*0.5333*
NoyolaBautistaJoel_2014	*0.9*	*1.1667*	*1*	*0.3*	*1*	*0.4615*
OcampoPolitoRigoberto_2010	*0.71*	*0.9*	*0.8*	*0.4*	*0.8*	*0.5333*
OlveraOrtegaJorge_2003	*0.567*	*0.6667*	*0.667*	*0.4*	*0.67*	*0.5*
OropezaRodriguezJoseLuis_2005	*0.9*	*1.1667*	*1*	*0.3*	*1*	*0.4615*
PeÃ±aAyalaAlejandro_2007	*0.675*	*0.875*	*0.75*	*0.3*	*0.75*	*0.4286*
PerezLeonJaimeAlfonso_2009	*0.463*	*0.625*	*0.5*	*0.2*	*0.5*	*0.2857*
QuinteroPerezGuillermo_2013	*0.888*	*1.125*	*1*	*0.4*	*1*	*0.5714*
QuinteroTellezRolando_2007	*0.95*	*1.3333*	*1*	*0.3*	*1*	*0.4615*
RenteriaAgualimpiaWalter_2009	*0.283*	*0.3333*	*0.333*	*0.1*	*0.33*	*0.1538*
ReyesMartinezMiguelDaniel_2015	*0.95*	*1.3333*	*1*	*0.3*	*1*	*0.4615*
ReyesTorresJesusJavier_2017	*0.71*	*0.9*	*0.8*	*0.4*	*0.8*	*0.5333*
RiveraLozaGabriela_2003	*0.617*	*0.8333*	*0.667*	*0.2*	*0.67*	*0.3077*
RodriguezMartinezMiguel_2011	*0.9*	*1.1667*	*1*	*0.3*	*1*	*0.4615*
RodriguezRomeroRaymundo_2020	*0.74*	*1*	*0.8*	*0.4*	*0.8*	*0.5333*
RojoRuizArturo_2008	*0.486*	*0.5714*	*0.571*	*0.4*	*0.57*	*0.4706*
RomeroXimilJoseManuel_2002	*0.667*	*1*	*0.667*	*0.2*	*0.67*	*0.3077*
SanchezFragaRodolfo_2013	*0.617*	*0.8333*	*0.667*	*0.2*	*0.67*	*0.3077*
SanchezFragaRodolfo_2016	*0.68*	*0.8*	*0.8*	*0.4*	*0.8*	*0.5333*
SuarezOropezaMiguel_2012	*0.713*	*1*	*0.75*	*0.3*	*0.75*	*0.4286*
TorresCruzNoe_2019	*0.71*	*0.9*	*0.8*	*0.4*	*0.8*	*0.5333*
TrejoSotoGloriaIrene_2002	*0.71*	*0.9*	*0.8*	*0.4*	*0.8*	*0.5333*
UriarteArciaAbrilValeria_2012	*0.9*	*1.1667*	*1*	*0.3*	*1*	*0.4615*
UriarteArciaAbrilValeria_2016	*0.283*	*0.3333*	*0.333*	*0.1*	*0.33*	*0.1538*

ValleChavezAbel_2012	*0.713*	*1*	*0.75*	*0.3*	*0.75*	*0.4286*
VanegasSanchezTonatiuhDaniel_2018	*0.9*	*1.1667*	*1*	*0.3*	*1*	*0.4615*
VarelaGarciaFrancisco_2002	*0.5*	*0.75*	*0.5*	*0.2*	*0.5*	*0.2857*
VazquezOropezaJonathan_2016	*0.57*	*0.8*	*0.6*	*0.3*	*0.6*	*0.4*
VelazquezAlcantaraRodrigoAlejandro_2017	*0.65*	*0.8571*	*0.714*	*0.5*	*0.71*	*0.5882*
VerasteguiBarrancoKarina_2007	*1*	*1.5*	*1*	*0.3*	*1*	*0.4615*
ZarateEscobedoRicardo_2018	*0.9*	*1.1667*	*1*	*0.3*	*1*	*0.4615*
ZhilaAlisa_2014	*0.888*	*1.125*	*1*	*0.4*	*1*	*0.5714*

Como se observa en la Tabla 5-6, los valores de la métrica *recall* y el modo propuesto "Onehot" son los mismos, sin embargo, *recall* no puede mostrar cuando el resultado a acertado con la "importancia" de la clasificación en la tesis (Alta, Media, Baja) de ahí que surja la necesidad de realizar modos de evaluación que tomen en consideración este factor, y lo evalúe ya sea penalizándolo o beneficiándolo. Estas consideraciones permiten visualizar la situación con la tesis: "CoronaBermudezErendira_2020" que tiene un valor en "Onehot" de 1, y en "Strict" de 0.97, reconociendo que existió una correcta identificación de las clasificaciones, no obstante fallo en la asignación de la etiqueta "importancia".

Con respecto a los valores bajos de *precision,* son debido a que toma en consideración los "Falsos Positivos", y dado que el sistema amplía la clasificación con respecto al *Golden Standard* (da una lista de posibles clasificaciones), el valor de *precision* se ve mermado. Esto se visualiza cuando se observa la matriz de confusión (obtenida con la función *multilabel_confusion_matrix* de la herramienta sci-kit learn) por cada tesis, como lo muestra la Ilustración 5-7, teniendo como estructura:

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

Siendo:

TN: *True Negatives* (Verdaderos Negativos)

FP: *False Positive* (Falsos Positivos)

FN: *False Negatives* (Falsos Negativos)

TP: *True Positive* (Verdaderos Positivos)

```
Confusion matrix for sample:  RomeroXimilJoseManuel_2002
[[2102    8]
 [   1    2]]
Confusion matrix for sample:  SanchezFragaRodolfo_2013
[[2102    8]
 [   1    2]]
Confusion matrix for sample:  SanchezFragaRodolfo_2016
[[2102    6]
 [   1    4]]
Confusion matrix for sample:  SuarezOropezaMiguel_2012
[[2102    7]
 [   1    3]]
Confusion matrix for sample:  TorresCruzNoe_2019
[[2102    6]
 [   1    4]]
Confusion matrix for sample:  TrejoSotoGloriaIrene_2002
[[2102    6]
 [   1    4]]
Confusion matrix for sample:  UriarteArciaAbrilValeria_2012
[[2103    7]
 [   0    3]]
```

Ilustración 5-7. Matrices de confusión para tesis.

De igual manera que hay matrices de confusión para cada tesis, las hay también para cada clasificación (véase anexos C y D para la información completa).

El reporte de clasificación de la herramienta sci-kit learn (función *classification_report*) permite visualizar las métricas *precision*, *recall* y *F1* por cada clasificación, tal y como lo muestra la Tabla 5-7 teniendo la posibilidad de observar las 213 clasificaciones únicas en el Golden Standard, las cuales se reconocieron en un 74.64% apropiadamente, es decir que al menos hubo una coincidencia con el *Golden Standard*.

Además de la descripción por cada clasificación se tienen diferentes tipos de promedios:

- *Micro avg:* Se hace un cómputo de *F1* donde se consideran el total de Verdaderos positivos (TP), Falsos negativos (FN) y Falsos positivos (FP). No importa la predicción para cada etiqueta en el conjunto.
- *Macro avg:* Se hace un cómputo de *F1* para cada etiqueta, y retorna el promedio sin considerar la proporción de cada etiqueta en el conjunto.
- *Weighted avg:* Se hace un cómputo de *F1* para cada etiqueta, y retorna el promedio considerando la proporción de cada etiqueta en el conjunto.
- *Samples avg:* Se hace un cómputo de *F1* para cada instancia, y retorna el promedio.

La descripción de cada una permitirá tomar en consideración la información relevante que aporta cada una a la hora de evaluar nuestro sistema, siendo que si se trabajó con un *dataset* desbalanceado donde todas las clases son igual de importantes, es preferible usar *macro avg* ya que trata todas las clases por igual, sin embargo si uno tiene un *dataset* desbalanceado pero se busca tomar en consideración aquellas clases que están más presentes en el *dataset*, entonces *weighted avg* sería las más recomendable. Por último, si se busca tener una métrica sin tomar en consideración las clases, *micro avg* sería la elección. Siendo *weighted avg* la más idónea para nuestro sistema.

Samples avg tiene como valores los mismos que en la Tabla 5-5, al ser analizadas por cada tesis (instancias).

Tabla 5-7. Reporte de clasificación de la herramienta sci-kit learn.

Hierarchy	*precision*	*recall*	*f1-score*	*support*
2. hardware	0.1	1	0.1818	1
2.2. communication hardware, interfaces and storage	0.1429	1	0.25	1
2.2.1. signal processing systems	0.5	1	0.6667	3
2.2.1.1. digital signal processing	0.5455	0.8571	0.6667	7
2.2.2. sensors and actuators	0.7	1	0.8235	7
2.2.4. displays and imagers	0	0	0	1
2.2.8. sensor applications and deployments	1	0.3333	0.5	3
2.2.9. sensor devices and platforms	1	0.5	0.6667	2
2.2.10. sound-based input / output	0	0	0	1
2.2.15. electro-mechanical devices	1	1	1	2
2.3. integrated circuits	0.1429	1	0.25	1
2.3.4.1. transistors	0.3333	1	0.5	1
2.3.5. logic circuits	1	1	1	1
2.3.6. reconfigurable logic and fpgas	0.75	1	0.8571	3
2.3.6.3. programmable logic elements	0.6	1	0.75	3
2.3.6.5. reconfigurable logic applications	1	1	1	1
2.6.1.2. hardware-software codesign	0	0	0	1
2.6.2. hardware description languages and compilation	0.5	1	0.6667	1
2.7.1.5. simulation and emulation	0	0	0	4
2.7.3.1. bug detection, localization and diagnosis	1	1	1	1
2.8.9. testing with distributed and parallel systems	0	0	0	1
2.9.3.5. signal integrity and noise analysis	0.6667	0.6667	0.6667	3
2.10.4.1. microelectromechanical systems	0.5	1	0.6667	1
2.10.10.4. quantum dots and cellular automata	1	1	1	7
3.1.2. parallel architectures	1	1	1	2
3.1.3. distributed architectures	0.4	1	0.5714	2
3.1.4.1. neural networks	0	0	0	1

3.1.4.2. reconfigurable computing	0.25	1	0.4	1
3.1.4.8. quantum computing	1	1	1	1
3.2.2. robotics	0.1667	1	0.2857	1
3.2.2.1. robotic components	0.6667	1	0.8	2
3.2.2.2. robotic control	0.6667	0.8	0.7273	5
3.2.3. sensors and actuators	0.1	1	0.1818	1
3.2.4. system on a chip	0	0	0	1
4.2. network protocols	0.2308	1	0.375	3
4.2.3. link-layer protocols	0.5	1	0.6667	1
4.2.4.1. routing protocols	0.5	1	0.6667	2
4.2.6. session protocols	0.5	1	0.6667	1
4.4.1. data path algorithms	0.5	1	0.6667	1
4.4.1.3. packet scheduling	0.3333	1	0.5	1
4.6.1. network security	1	1	1	1
4.6.1.1. security protocols	1	1	1	1
4.6.1.2. web protocol security	0	0	0	1
4.6.5. network reliability	0	0	0	1
4.8.6.1. sensor networks	1	1	1	1
4.8.8.1. peer-to-peer networks	1	1	1	1
4.8.9. wireless access networks	1	1	1	1
4.8.10. ad hoc networks	0.5	1	0.6667	2
4.8.10.1. mobile ad hoc networks	0.6667	1	0.8	2
5.1.1.2.3. virtual machines	1	1	1	1
5.1.1.3. operating systems	0.1667	1	0.2857	1
5.1.1.3.2. memory management	1	0.5	0.6667	2
5.1.2.5.2. client-server architectures	0	0	0	1
5.1.3.2.2. software verification	0	0	0	1
5.2.1.2. language features	0	0	0	1
5.2.3.10. parsers	0.6667	1	0.8	2
5.2.4.1.1. extensible markup language (xml)	0.3333	1	0.5	1
5.2.4.8. command and control languages	0	0	0	1
5.2.8. software libraries and repositories	0.3333	1	0.5	1
5.3.2.1. software development methods	1	1	1	2
5.3.3. software development techniques	0.3333	1	0.5	1
5.3.4. software verification and validation	0	0	0	1
6.1.6.1. parallel computing models	0	0	0	1
6.3. computational complexity and cryptography	1	1	1	2
6.5.2.1. scheduling algorithms	0	0	0	1
6.5.2.3. routing and network design problems	0	0	0	1
6.5.3. mathematical optimization	0.2	0.5	0.2857	2
6.5.3.1.2.2. evolutionary algorithms	1	1	1	1
6.5.3.2. continuous optimization	0	0	0	1
6.5.3.2.8. bio-inspired optimization	0.5	1	0.6667	1

6.5.3.3.3. bio-inspired optimization	1	1	1	1
6.6.4. random walks and markov chains	0.5	1	0.6667	1
6.7.1.4.1. support vector machines	1	1	1	1
6.7.1.6. bayesian analysis	0	0	0	1
6.7.3.6. database interoperability	0	0	0	1
6.7.3.9. data integration	1	0.5	0.6667	2
6.8. semantics and reasoning	0.6429	0.8182	0.72	11
7.1.2.9. network flows	0	0	0	1
7.2. probability and statistics	1	0.3333	0.5	3
7.2.3.5. markov-chain monte carlo methods	0	0	0	1
7.2.3.7. kalman filters and hidden markov models	0.75	1	0.8571	3
7.2.5. statistical paradigms	1	1	1	1
7.2.6. stochastic processes	1	1	1	1
7.4.1. coding theory	0	0	0	1
7.5.1.1. computation of transforms	0.6667	1	0.8	2
7.5.1.8. interval arithmetic	1	1	1	1
7.5.2.1.2.2. evolutionary algorithms	1	1	1	1
7.5.2.2.8. bio-inspired optimization	1	1	1	1
7.5.2.3.3. bio-inspired optimization	1	1	1	1
7.5.7. nonlinear equations	0	0	0	1
8.1.1. database design and models	0.5	1	0.6667	1
8.1.1.5.2. data streams	0	0	0	1
8.1.3.2. database query processing	0.3077	1	0.4706	4
8.1.3.5. parallel and distributed dbmss	1	1	1	1
8.1.4.2. xml query languages	0	0	0	1
8.1.4.2.1. xpath	0	0	0	1
8.1.6. information integration	1	1	1	1
8.1.6.2. extraction, transformation and loading	1	1	1	2
8.1.6.7. entity resolution	1	1	1	1
8.3. information systems applications	0.1429	0.5	0.2222	2
8.3.2. collaborative and social computing systems and tools	1	1	1	1
8.3.3. spatial-temporal systems	0.0833	1	0.1538	1
8.3.3.1. location based services	0.25	1	0.4	1
8.3.3.2. geographic information systems	1	0.9231	0.96	13
8.3.3.4. data streaming	0	0	0	1
8.3.7.2. multimedia streaming	0.5	1	0.6667	1
8.3.8.4. clustering	0.1429	1	0.25	1
8.3.8.6. data stream mining	0	0	0	1
8.3.9. digital libraries and archives	1	1	1	2
8.4.3.2. data extraction and integration	1	1	1	2
8.4.3.4. traffic analysis	0	0	0	1
8.4.6. web services	0.3333	1	0.5	1
8.4.7.1. semantic web description languages	0.625	1	0.7692	5

8.4.7.1.1. resource description framework (rdf)	0.25	1	0.4	1
8.4.7.2. markup languages	1	1	1	1
8.4.7.2.1. extensible markup language (xml)	0.6667	1	0.8	2
8.5. information retrieval	0.3333	1	0.5	2
8.5.1.6. ontologies	0.6364	1	0.7778	7
8.5.1.7. dictionaries	0.25	0.5	0.3333	2
8.5.4.2. probabilistic retrieval models	0	0	0	1
8.5.4.4. similarity measures	0.6	1	0.75	3
8.5.5. retrieval tasks and goals	0.1	1	0.1818	1
8.5.5.2. document filtering	0	0	0	1
8.5.5.4. information extraction	0.6667	1	0.8	2
8.5.5.5. sentiment analysis	0	0	0	1
8.5.5.8. clustering and classification	0.3871	1	0.5581	12
8.5.8.1.1. structured text search	0	0	0	1
9.1. cryptography	1	1	1	3
9.1.2. symmetric cryptography and hash functions	0.3333	1	0.5	1
9.3.1. authentication	1	1	1	2
9.4.1. malware and its mitigation	1	1	1	3
9.7.1. security protocols	1	1	1	1
10.1.2.5. virtual reality	1	1	1	1
10.1.3. interaction devices	1	1	1	1
10.1.3.4. keyboards	0.3333	1	0.5	1
10.1.3.6. touch screens	1	1	1	1
10.2.1.1. user interface design	0	0	0	1
10.4.3. ubiquitous and mobile devices	0.2	1	0.3333	1
10.4.3.1. smartphones	1	1	1	1
10.4.3.3. mobile phones	0	0	0	1
10.4.3.4. mobile devices	0.5	1	0.6667	1
10.5. visualization	0.5	1	0.6667	1
10.5.2.3. geographic visualization	0	0	0	1
10.5.6. visualization design and evaluation methods	1	1	1	1
11.2. parallel computing methodologies	0	0	0	1
11.2.1.3. shared memory algorithms	0	0	0	2
11.2.2. parallel programming languages	0.5	0.5	0.5	2
11.3. artificial intelligence	0.05	1	0.0952	1
11.3.1. natural language processing	0.4231	1	0.5946	11
11.3.1.3. discourse, dialogue and pragmatics	0	0	0	1
11.3.1.4. natural language generation	1	1	1	2
11.3.1.5. speech recognition	0.75	1	0.8571	3
11.3.1.6. lexical semantics	0.5	0.6667	0.5714	3
11.3.1.7. phonology / morphology	0.3333	1	0.5	1
11.3.1.8. language resources	1	0.5	0.6667	2
11.3.2. knowledge representation and reasoning	0.1667	1	0.2857	1

11.3.2.2. semantic networks	1	1	1	1
11.3.2.5. vagueness and fuzzy logic	0.5	1	0.6667	5
11.3.2.9. ontology engineering	0.25	1	0.4	1
11.3.4.1. heuristic function construction	0	0	0	1
11.3.5.2. computational control theory	0	0	0	1
11.3.5.3. motion path planning	0	0	0	1
11.3.8. computer vision	0.75	1	0.8571	3
11.3.8.1. computer vision tasks	0.3333	1	0.5	1
11.3.8.1.5. visual content-based indexing and retrieval	1	1	1	1
11.3.8.1.7. vision for robotics	0	0	0	3
11.3.8.2.6. 3d imaging	1	1	1	1
11.3.8.3.1. image representations	0	0	0	1
11.3.8.4.2. image segmentation	1	0.6667	0.8	3
11.3.8.4.9. reconstruction	0	0	0	1
11.4. machine learning	0.2857	1	0.4444	4
11.4.1. learning paradigms	0.5	0.8333	0.625	6
11.4.1.1. supervised learning	0.6	1	0.75	12
11.4.1.1.1. ranking	0	0	0	1
11.4.1.1.3. supervised learning by classification	0.4	0.6667	0.5	3
11.4.1.2. unsupervised learning	0.375	1	0.5455	3
11.4.3. machine learning approaches	0.0833	1	0.1538	1
11.4.3.2.1. support vector machines	0.5	1	0.6667	3
11.4.3.3. neural networks	0.6842	1	0.8125	13
11.4.3.4.1. inductive logic learning	1	1	1	1
11.4.3.5.5. latent variable models	1	1	1	1
11.4.3.5.6. bayesian network models	1	1	1	1
11.4.3.6. learning linear models	0.2	1	0.3333	1
11.4.3.6.1. perceptron algorithm	0.25	0.5	0.3333	2
11.4.3.8. rule learning	0	0	0	1
11.4.3.11. partially-observable markov decision processes	0	0	0	1
11.4.3.14. bio-inspired approaches	1	1	1	1
11.4.4. machine learning algorithms	0.5	1	0.6667	1
11.4.4.3. spectral methods	0	0	0	1
11.5. modeling and simulation	0.4545	0.8333	0.5882	6
11.5.1. model development and analysis	0.5	0.3333	0.4	3
11.5.3. simulation types and techniques	0.3333	0.6667	0.4444	3
11.5.3.5. discrete-event simulation	0	0	0	1
11.5.3.6. agent / discrete models	0.8	1	0.8889	8
11.5.3.13. massively parallel and high-performance simulations	0	0	0	1
11.6.3.2. image processing	0.5	1	0.6667	1
12.2.8. enterprise ontologies, taxonomies and vocabularies	0.1667	1	0.2857	2
12.2.15.2. information integration and interoperability	1	1	1	1

12.3.1.1. avionics	0.5	1	0.6667	2
12.3.5. earth and atmospheric sciences	0.5	1	0.6667	2
12.3.5.1. environmental sciences	0.3333	1	0.5	1
12.3.9. electronics	0.2	1	0.3333	1
12.4. life and medical sciences	0.5	1	0.6667	3
12.4.6. health informatics	0	0	0	2
12.4.9.2. proteomics	0.5	1	0.6667	1
12.5.4. economics	0.5	1	0.6667	1
12.8.2. publishing	1	1	1	2
12.9.5. transportation	0.5	1	0.6667	3
12.10. education	0	0	0	1
13.1.2.1.5. systems development	0.5	1	0.6667	1
13.1.2.3. software management	1	1	1	1
13.2.4. surveillance	1	1	1	2
13.2.7.4. malware / spyware crime	0.3333	1	0.5	1
micro avg	0.334	0.7915	0.4698	422
macro avg	0.0484	0.0709	0.0534	422
weighted avg	0.5503	0.7915	0.6132	422
samples avg	0.334	0.8104	0.4666	422

5.1.4. *Evaluación y Visualización temporal*

Varias tablas se usaron para dar respuesta a las preguntas planteadas con respecto a la evaluación temporal del desarrollo de temáticas en el CIC, mientras que en otras ocasiones se visualiza de mejor manera en una gráfica.

5.1.4.1. ¿Cuáles clasificaciones del ACM desarrolladas en el CIC se les da más continuidad (se vuelven a usar)?

Para responder esta pregunta, se seleccionó 5 clasificaciones que estuvieran presentes en mayor cantidad, y que sean constantes a lo largo del tiempo, que muestra la Tabla 5-8.

Considerando que en nuestro análisis no hay tesis del 2001, se tiene en total 20 años analizados (2000-2020).

Tabla 5-8. Clasificaciones que se les da más continuidad.

Clasificación	*Frecuencia/Total Años*
11.3.1. Natural language processing	20/20
8.5.5.8. Clustering and classification	19/20
4.2. Network protocols	18/20
11.4.3.3. Neural networks	18/20
11.4.1.1. Supervised learning	18/20

5.1.4.2. ¿QUÉ CLASIFICACIONES DEL ACM DESARROLLADAS EN EL CIC TIENEN UNA FUERTE PRODUCCIÓN EN LOS ÚLTIMOS AÑOS?

Para responder esta pregunta podemos tomar los últimos 5 años y revisar las 5 clasificaciones que aparecen con más frecuencia, que muestra la Tabla 5-9.

Tabla 5-9. Clasificaciones con fuerte producción en 2016-2020.

Clasificación	*Frecuencia*
8.5.5.8 Clustering and classification	94
11.4.1.1. Supervised learning	81
11.4.3.3. Neural networks	54
11.3.1. Natural language processing	47
11.4. Machine Learning	45

5.1.4.3. ¿QUÉ CLASIFICACIONES DEL ACM DESARROLLADAS EN EL CIC ESTÁN EN EL OLVIDO?

Para responder esta pregunta, se seleccionó 5 clasificaciones de los primeros 10 años que estaban en gran manera presentes, y que después ya no están, que muestra la Tabla 5-10.

Tabla 5-10. Clasificaciones que están en el olvido.

Clasificación
5.1.3.1.4. Consistency
11.3.5.3. Motion path planning
8.1.3.9.2. Deadlocks
3.3. Real-time systems
2.9.3. Hardware reliability

5.1.4.4. ¿CUÁLES CLASIFICACIONES DEL ACM DESARROLLADAS EN EL CIC SON NUEVAS?

Para responder esta pregunta, se seleccionó 5 clasificaciones que solamente aparecen una vez. Entre más específicas (nodos de bajo nivel) mejor, que muestra la Tabla 5-11.

Tabla 5-11. Clasificaciones que son nuevas.

Clasificación
5.1.1.3.2.2. main memory
5.1.1.2.2.1. message oriented middleware

5.1.1.3.4.3. message passing
6.5.3.1.2.3. tabu search
5.1.1.3.2.6. secondary storage

5.1.4.5. ¿CUÁLES CLASIFICACIONES DEL ACM SON LAS QUE MÁS SE HAN DESARROLLADO POR AÑO EN EL CIC?

De la Ilustración 5-8 a la Ilustración 5-27 muestran el top 5 de clasificaciones por año, con el número de tesis analizadas en ese año (No. Theses) y la frecuencia de cada clasificación.

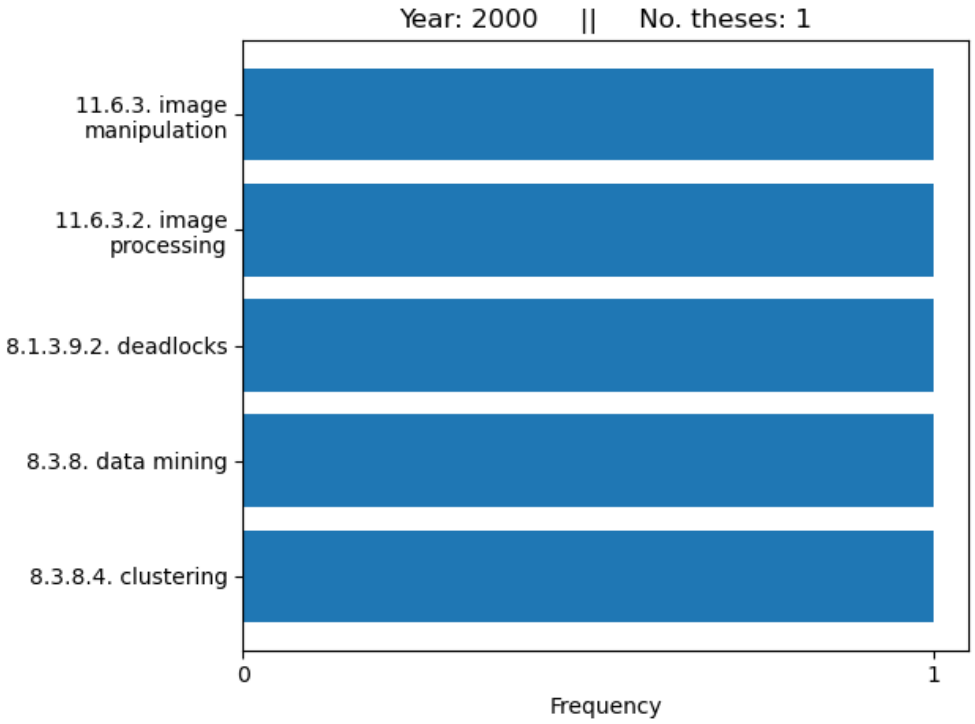


Ilustración 5-8. Top 5 clasificaciones del año 2000.

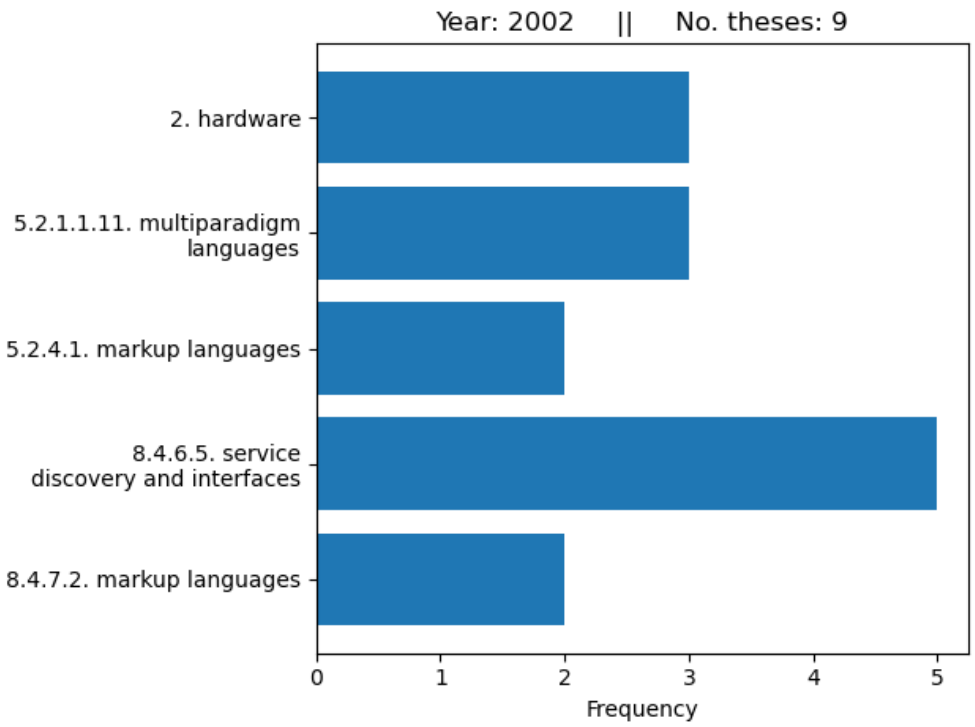


Ilustración 5-9. Top 5 clasificaciones del año 2002.

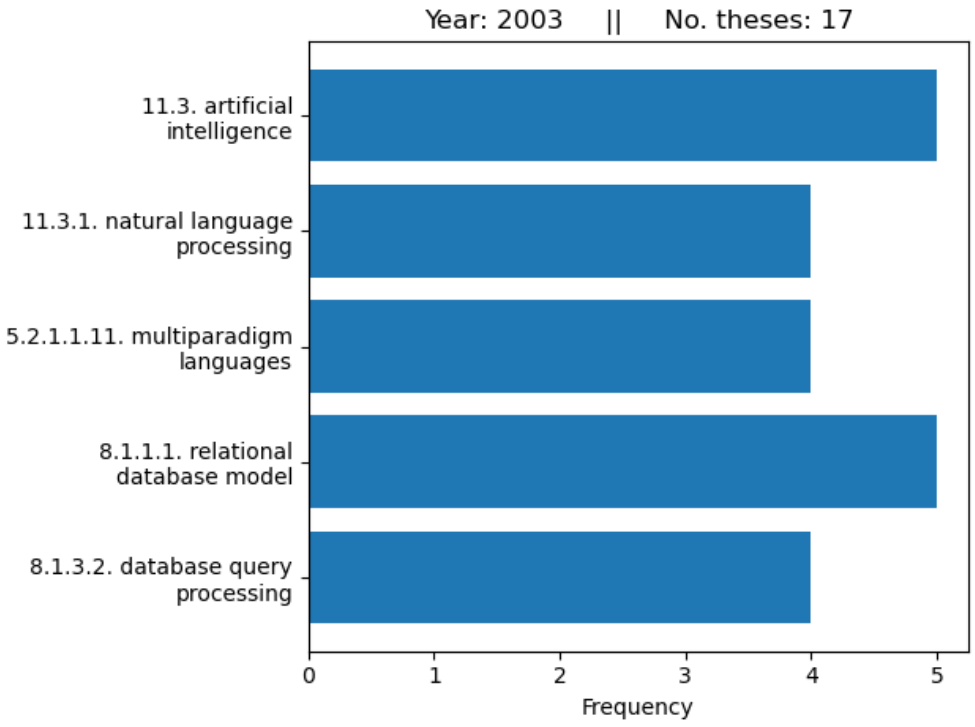


Ilustración 5-10. Top 5 clasificaciones del año 2003.

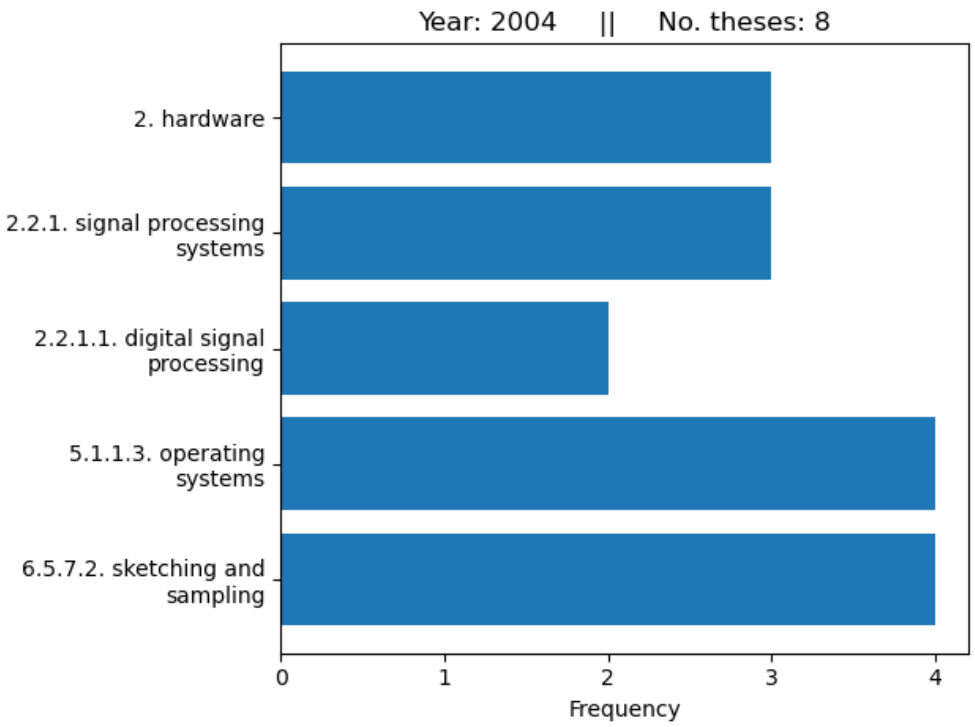


Ilustración 5-11. Top 5 clasificaciones del año 2004.

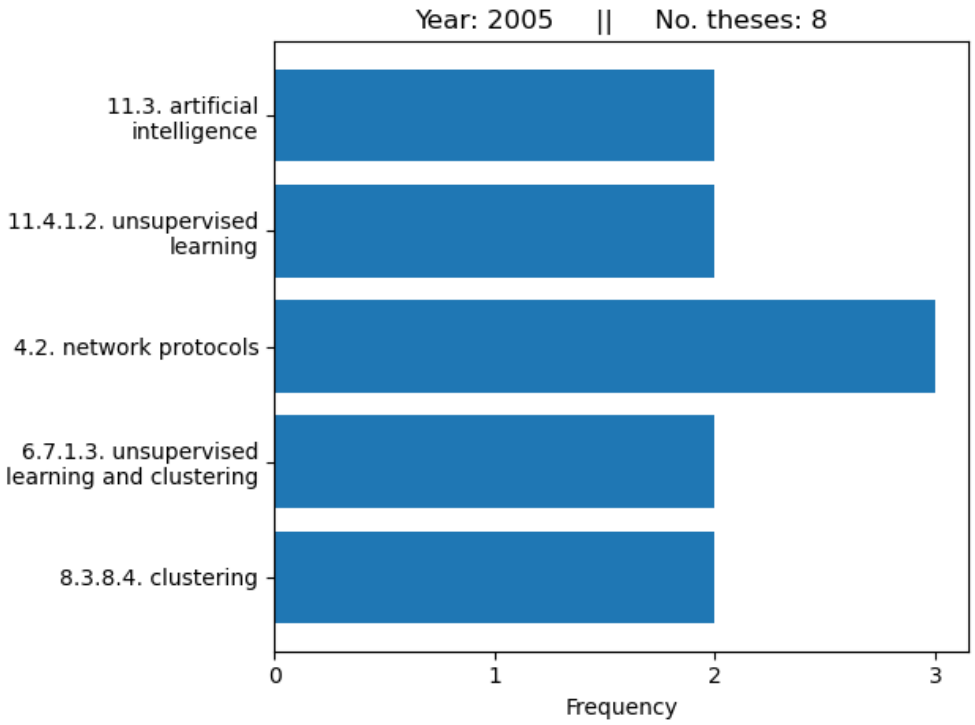


Ilustración 5-12. Top 5 clasificaciones del año 2005.

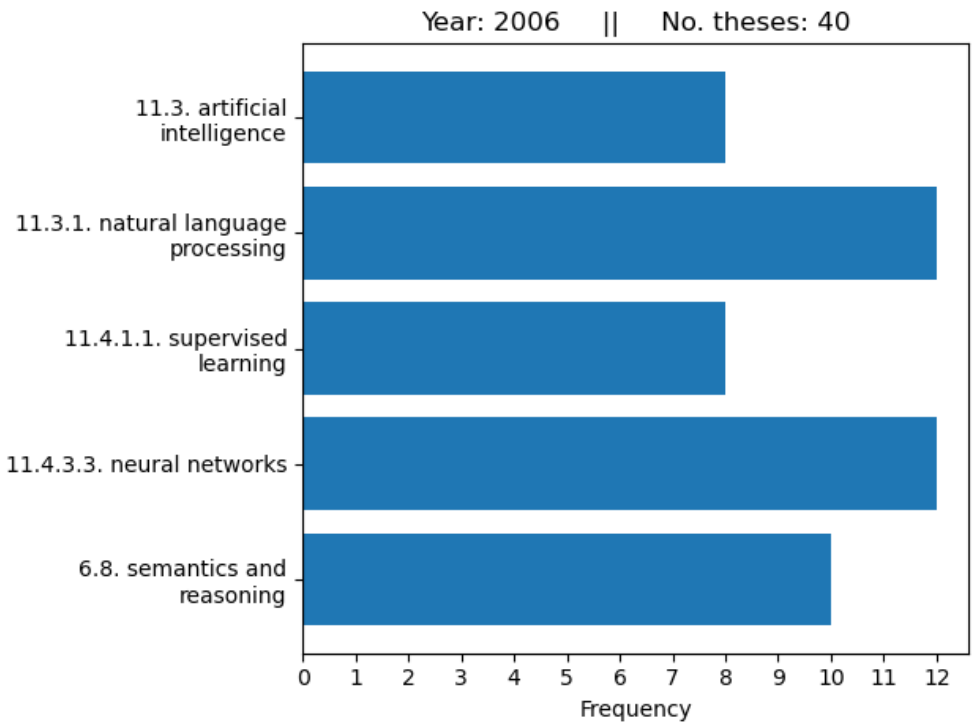


Ilustración 5-13. Top 5 clasificaciones del año 2006.

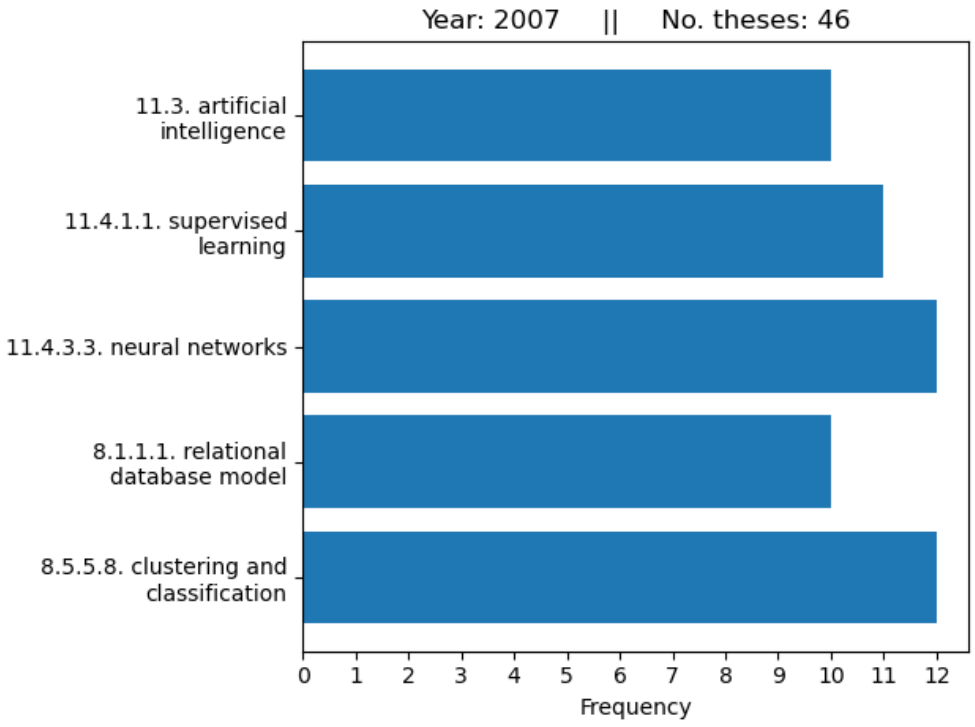


Ilustración 5-14. Top 5 clasificaciones del año 2007.

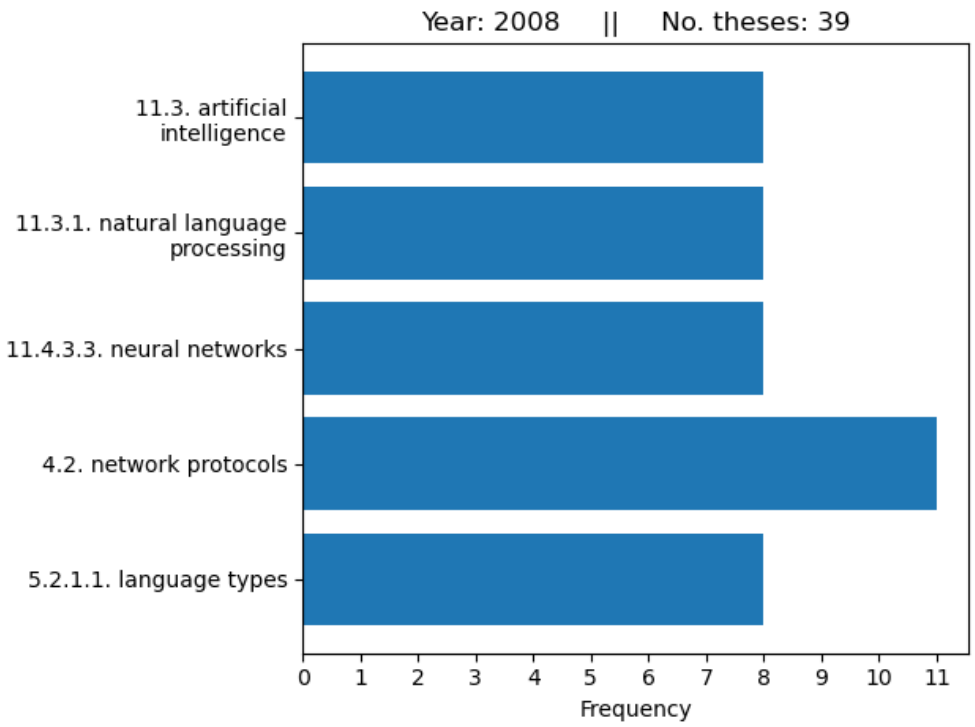


Ilustración 5-15. Top 5 clasificaciones del año 2008.

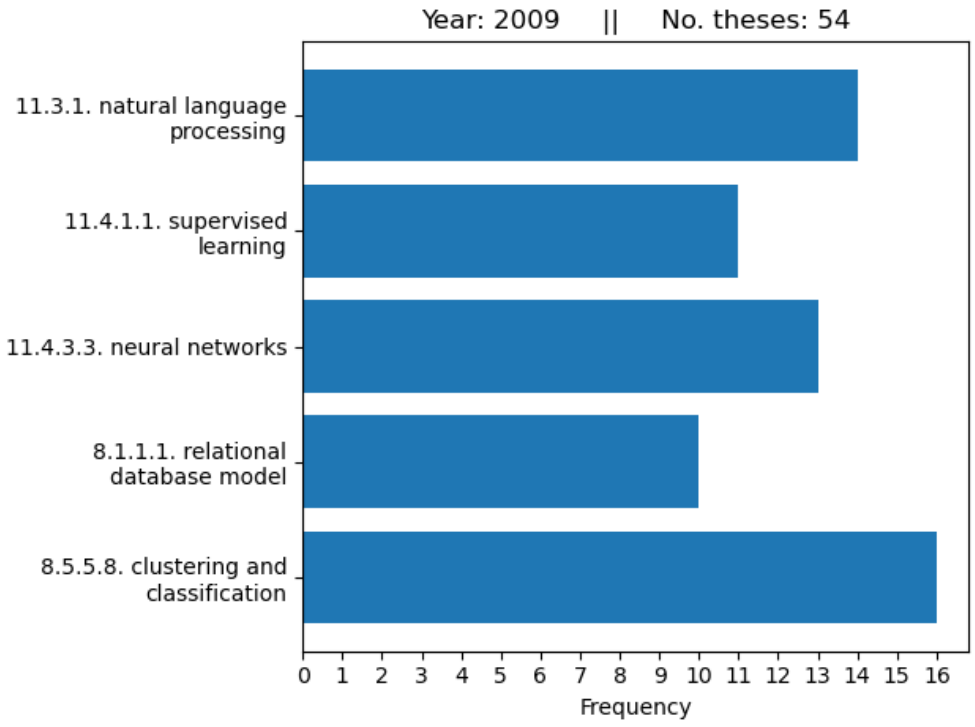


Ilustración 5-16. Top 5 clasificaciones del año 2009.

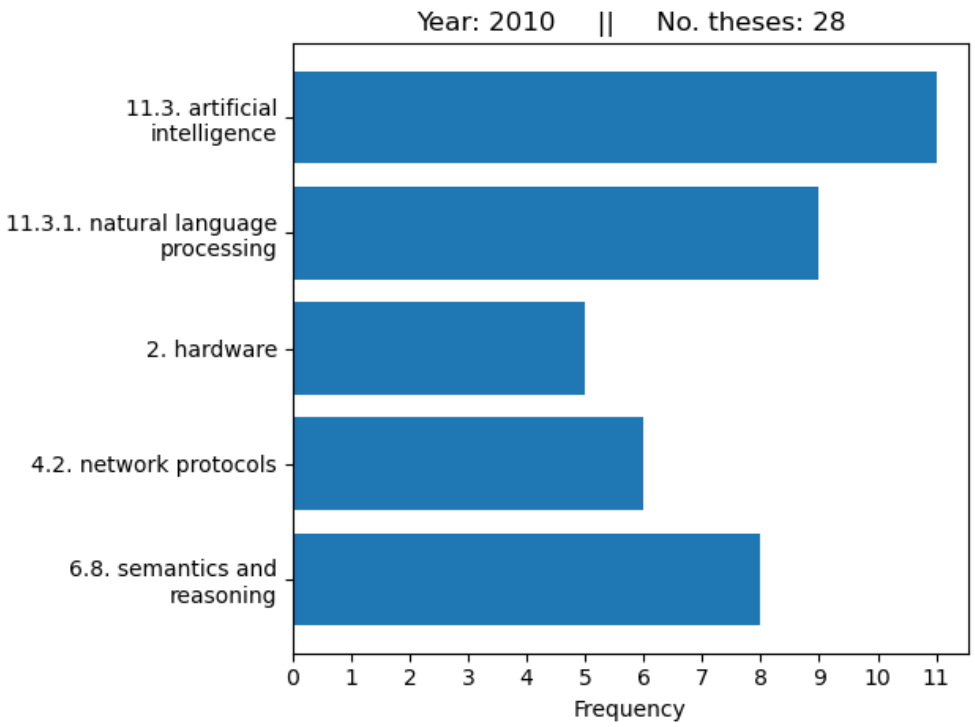


Ilustración 5-17. Top 5 clasificaciones del año 2010.

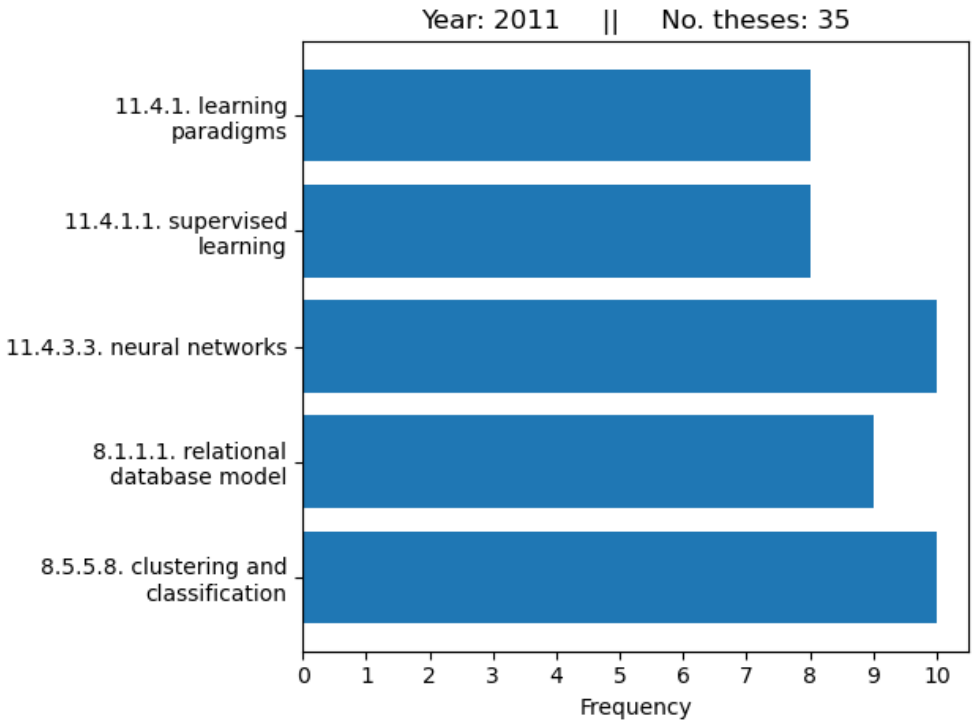


Ilustración 5-18. Top 5 clasificaciones del año 2011.

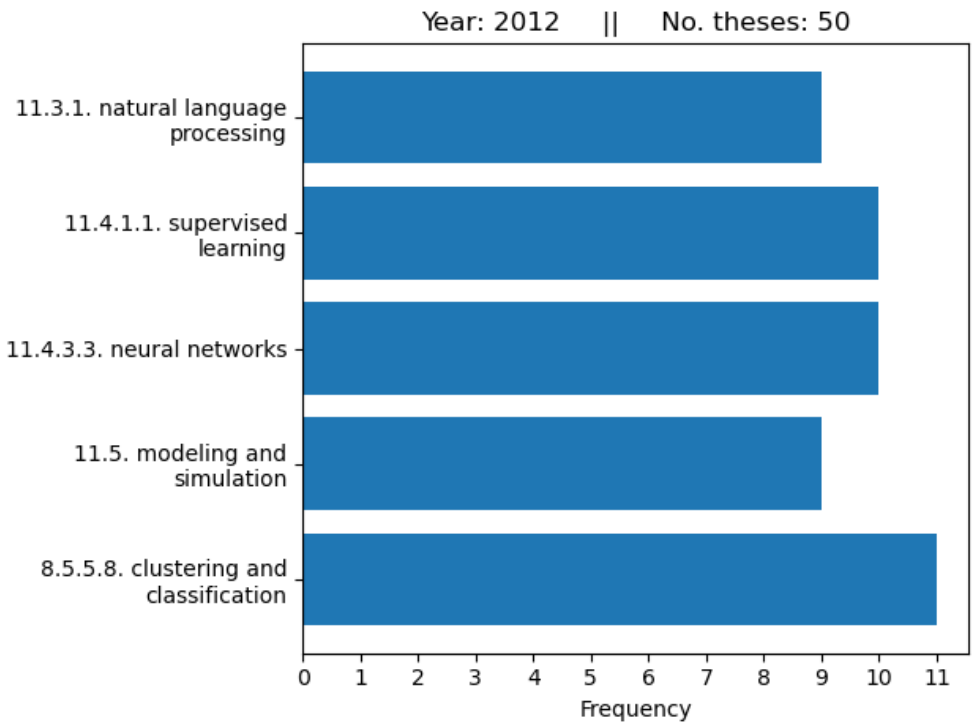


Ilustración 5-19. Top 5 clasificaciones del año 2012.

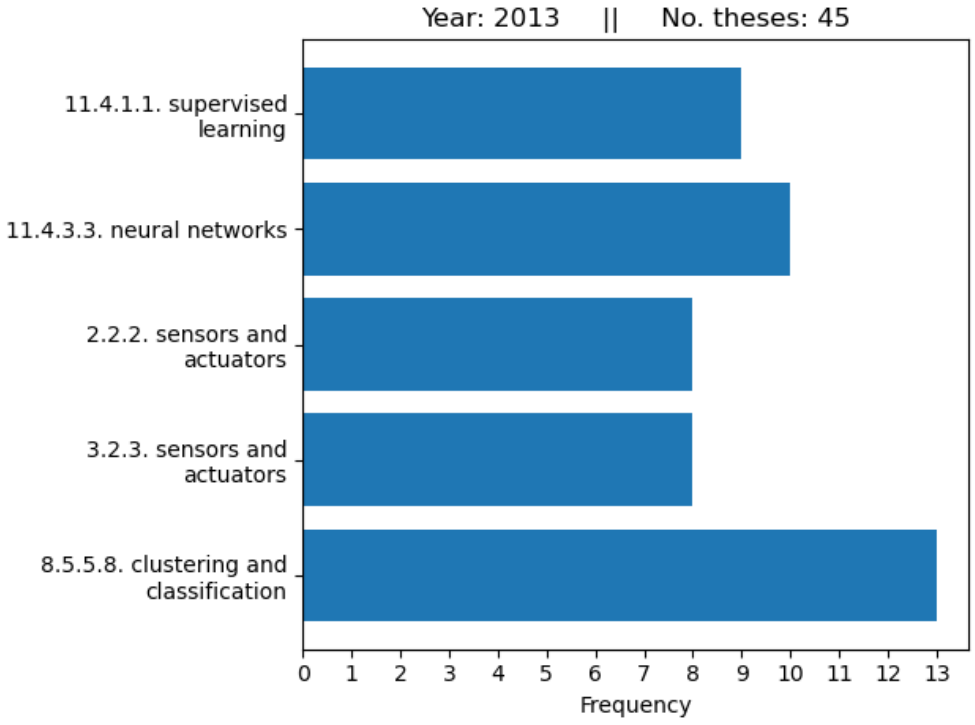


Ilustración 5-20. Top 5 clasificaciones del año 2013.

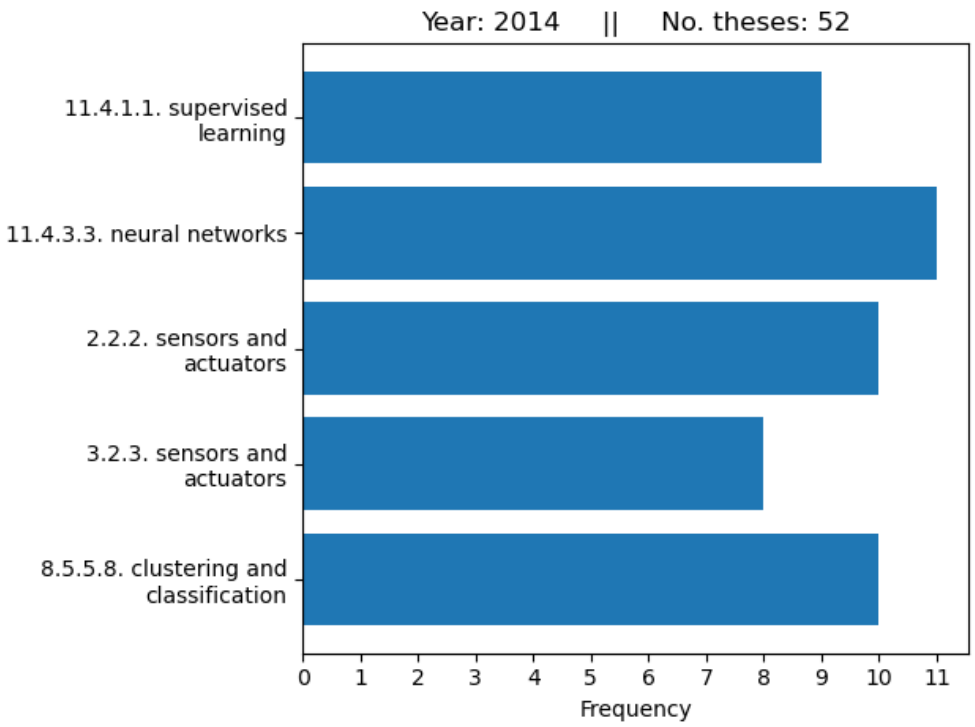


Ilustración 5-21. Top 5 clasificaciones del año 2014.

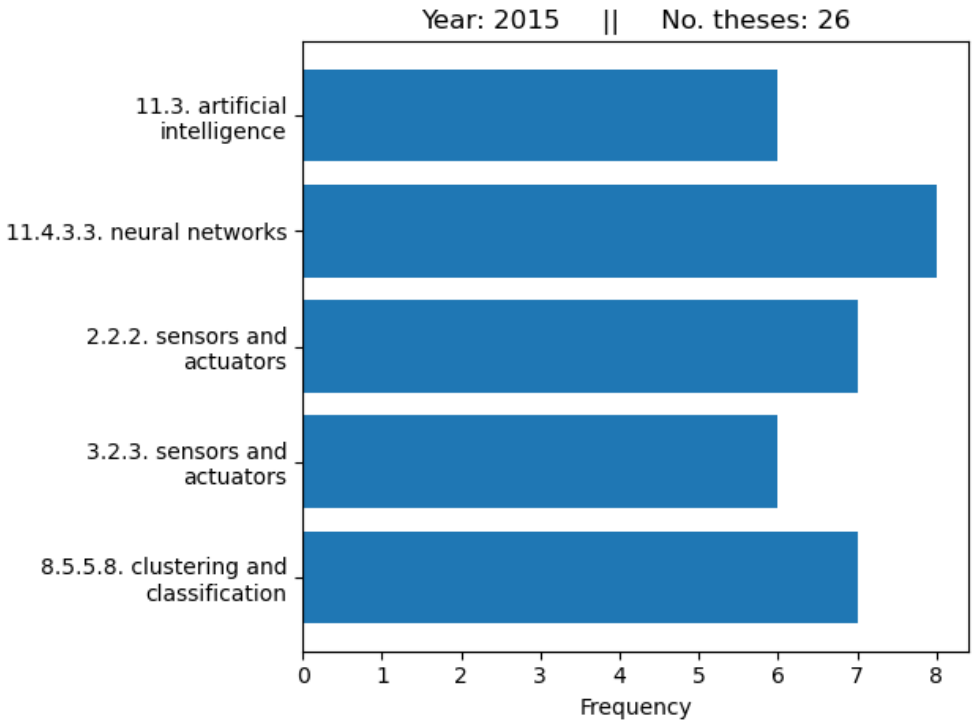


Ilustración 5-22. Top 5 clasificaciones del año 2015.

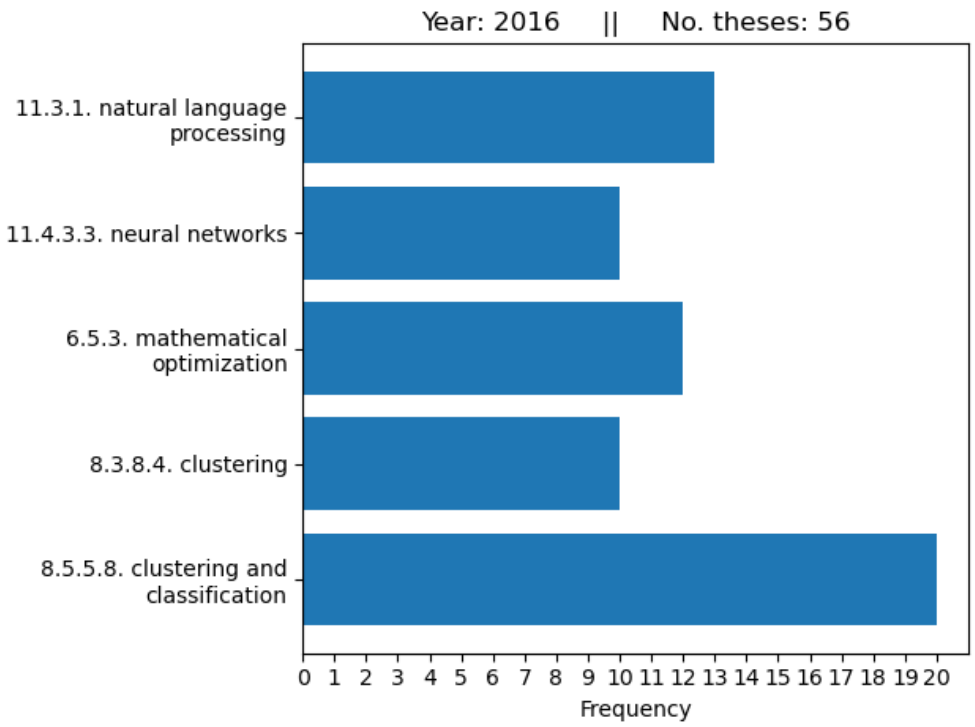


Ilustración 5-23. Top 5 clasificaciones del año 2016.

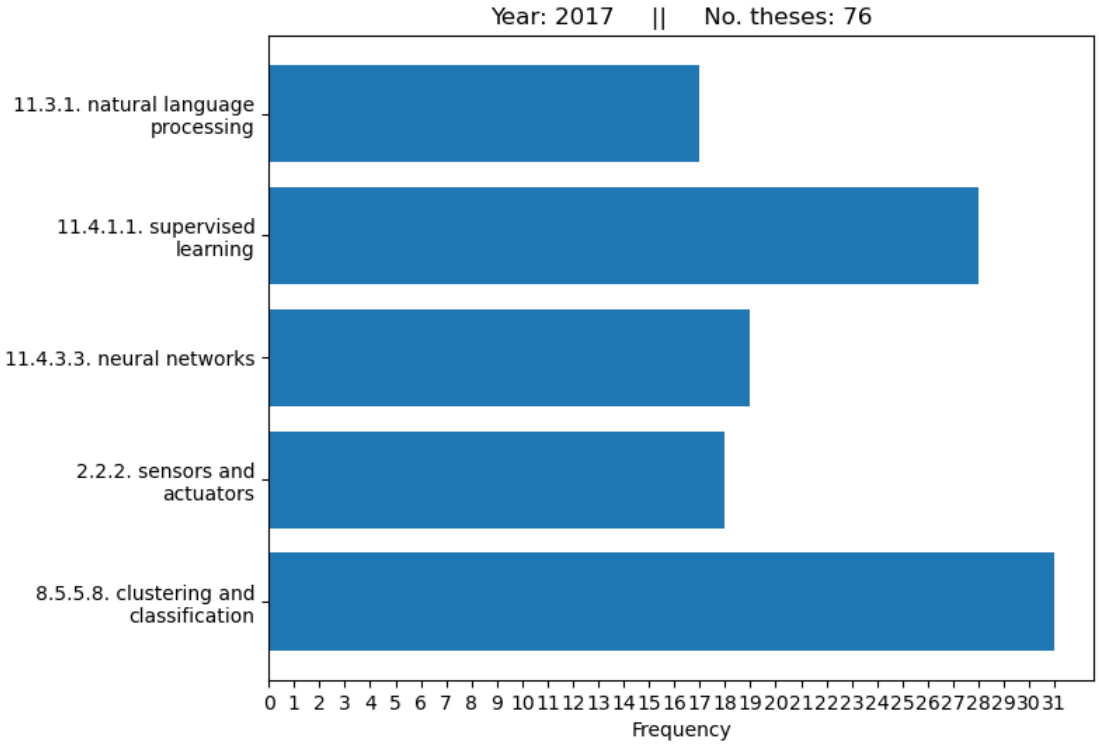


Ilustración 5-24. Top 5 clasificaciones del año 2017.

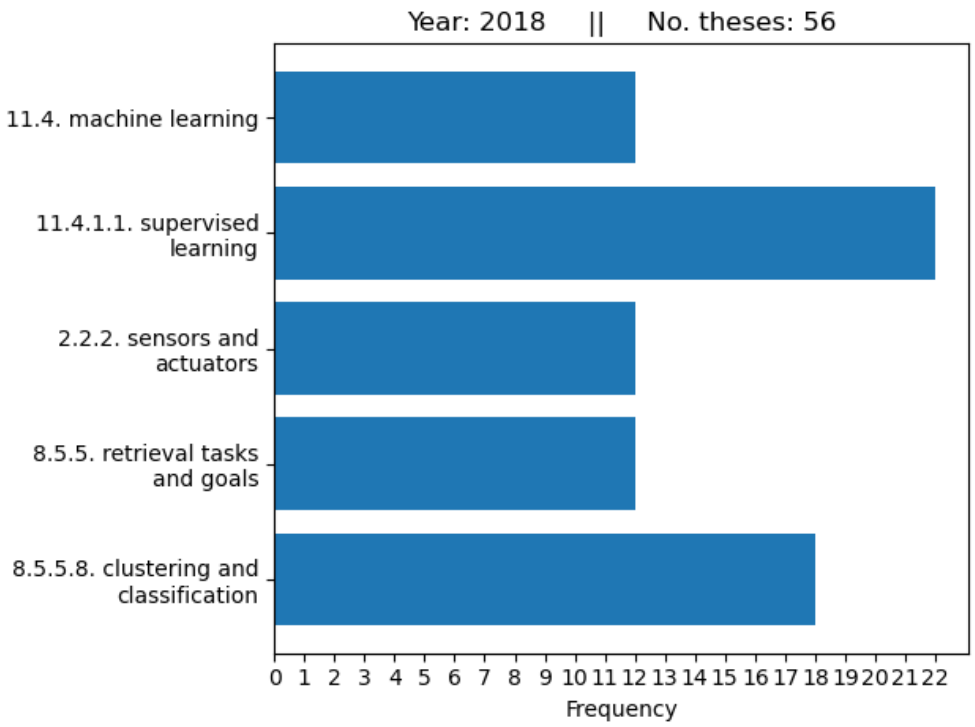


Ilustración 5-25. Top 5 clasificaciones del año 2018.

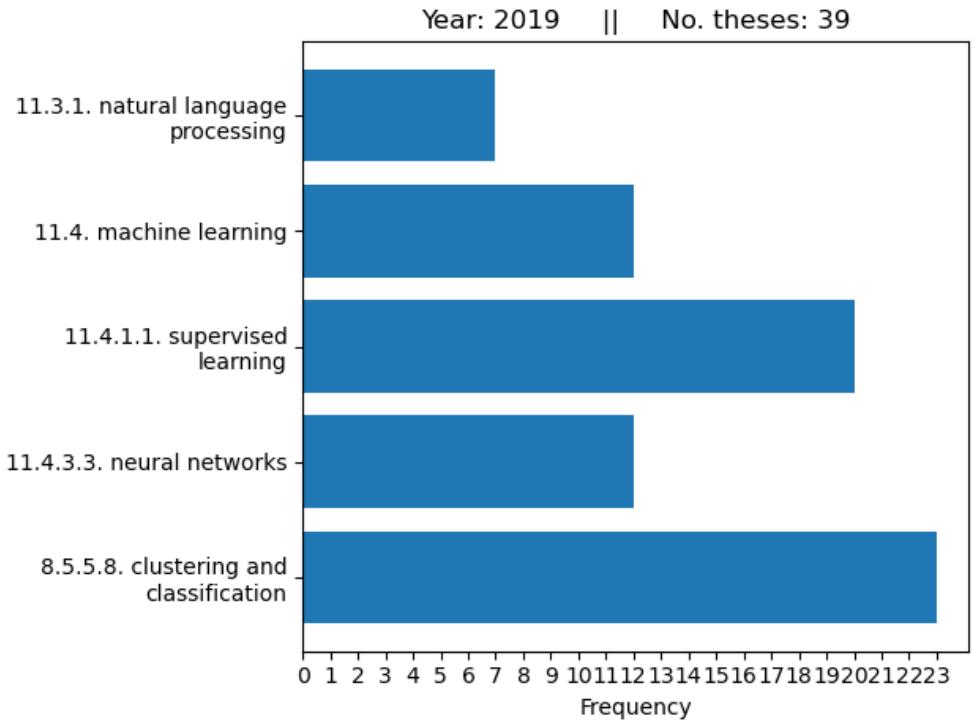


Ilustración 5-26. Top 5 clasificaciones del año 2019.

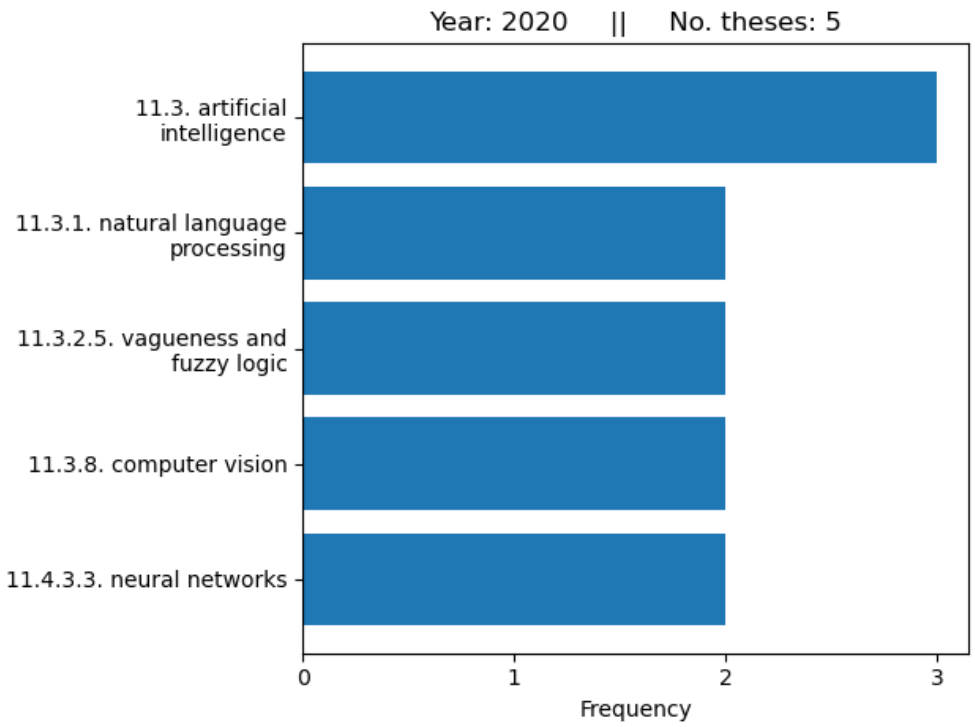


Ilustración 5-27. Top 5 clasificaciones del año 2020.

5.1.4.6. ¿CÓMO HA SIDO EL DESARROLLO A TRAVÉS DEL TIEMPO DE LAS CLASIFICACIONES DEL ACM HISTÓRICAMENTE MÁS USADAS EN EL CIC?

Las clasificaciones históricamente más usadas son:

- "11.4.3.3. Neural Networks"
- "8.1.3.2. Database Query Processing"
- “2.2.2. Sensors and Actuators”
- “8.5.5. Retrieval tasks and goals”
- "6.8. Semantics and Reasoning"

Y su desarrollo se puede visualizar con las gráficas que se muestran en Ilustración 5-28 hasta la Ilustración 5-32.

Las gráficas expresan en porcentaje, que tanto se desarrolló una clasificación comparándolas con el total de tesis desarrolladas por año. Se debe de tomar en cuenta que, por cada año, la cantidad de tesis analizada no es la misma.

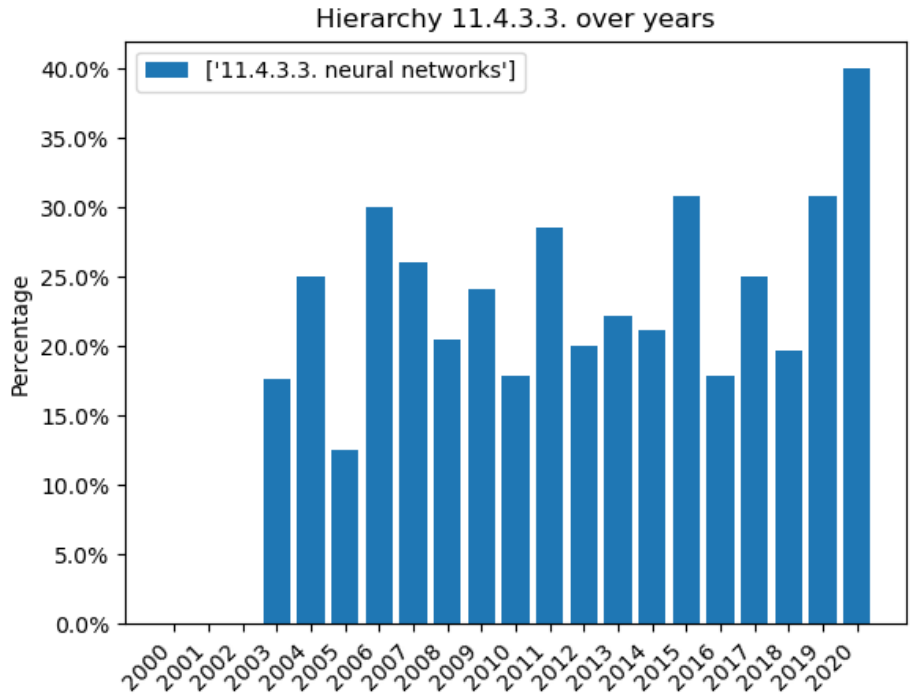


Ilustración 5-28. Clasificación "11.4.3.3. Neural Networks" a través del tiempo.

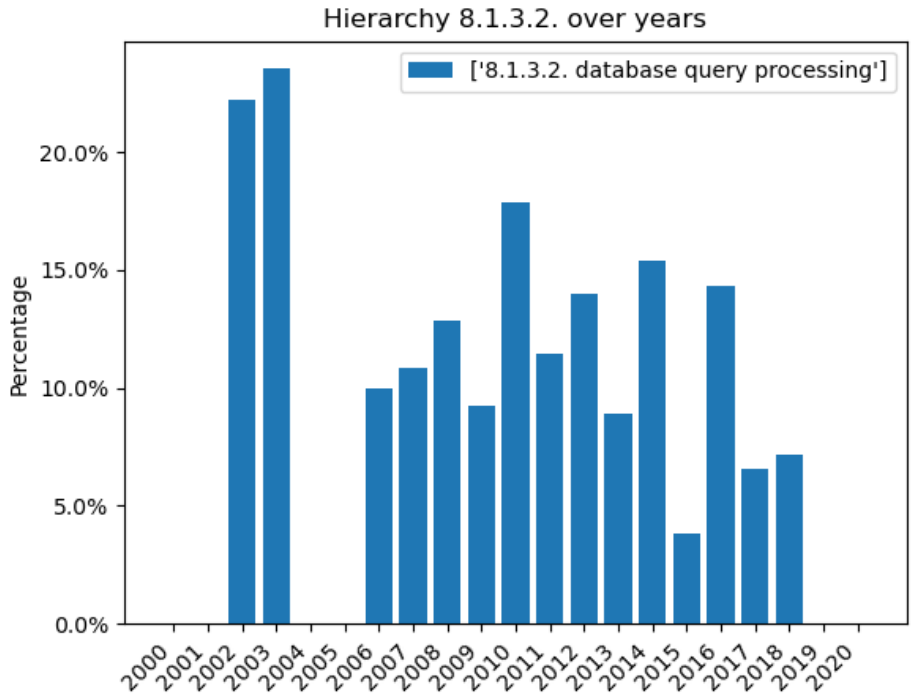


Ilustración 5-29. Clasificación "8.1.3.2. Database Query Processing" a través del tiempo.

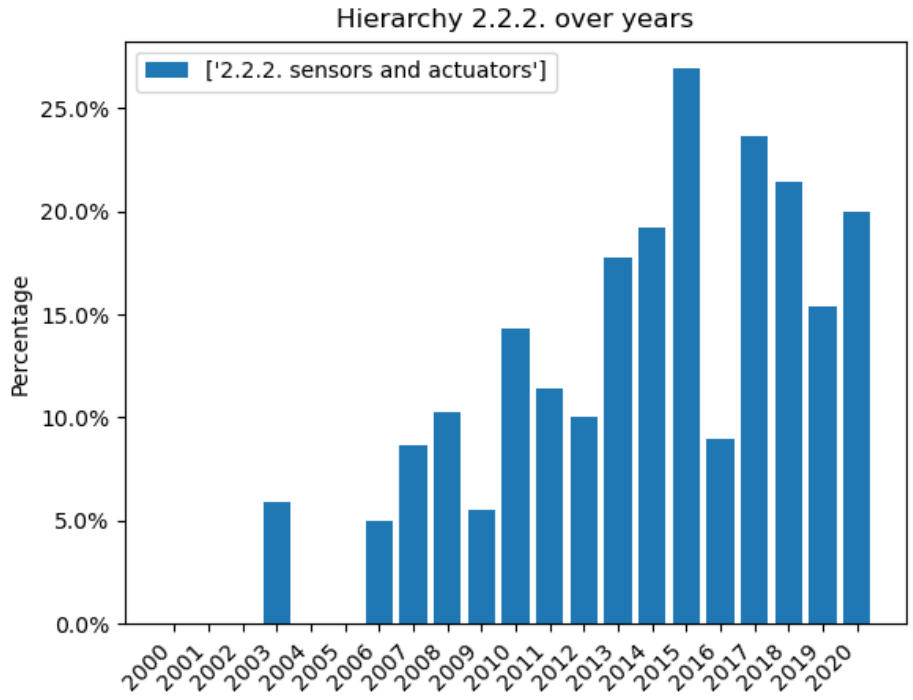


Ilustración 5-30. Clasificación "2.2.2. Sensors and actuators" a través del tiempo.

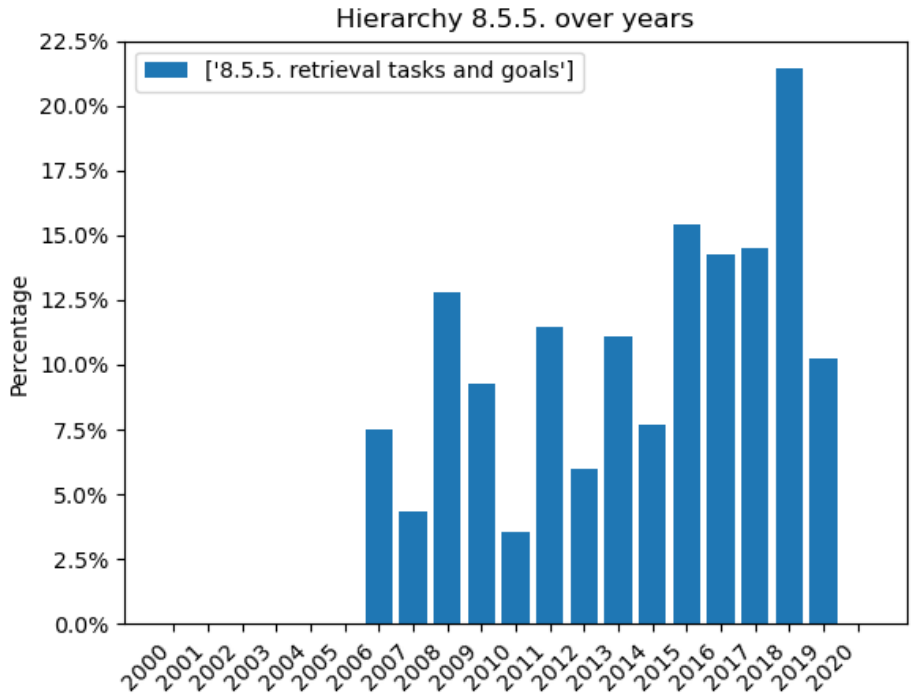


Ilustración 5-31. Clasificación "8.5.5. Retrieval tasks and goals" a través del tiempo.

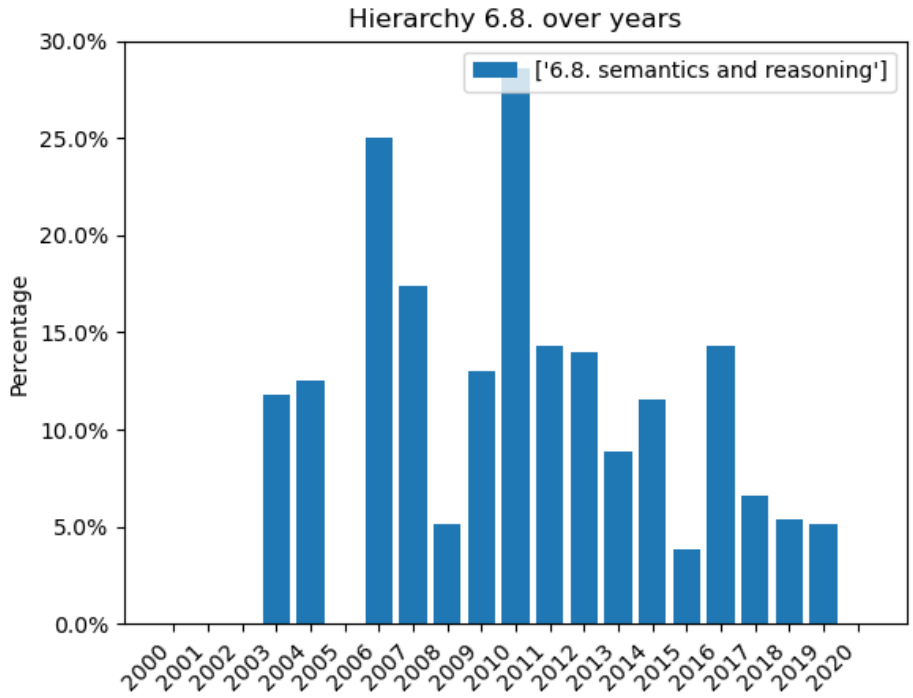


Ilustración 5-32. Clasificación "6.8. Semantics and reasoning" a través del tiempo.

Es evidente que unas tienen un comportamiento en aumento como lo puede ser la clasificación "2.2.2. Sensors and actuators", mientras que otras están presentes de forma consistente a través de los años como "11.4.3.3. Neural Networks".

5.1.4.7. Producción de tesis por año

Por último y no menos importante se puede observar la producción de tesis por año en la Tabla 5-12, sin embargo, esta información está supeditada a las plataformas que tiene, en este caso el IPN, para la publicación de tesis de manera digitalizada (repositorio Dspace[1] y repositorio digital IPN[2]) y que dependen de actualizaciones y otros factores para mantener actualizados los registro. Y en palabras de un encargado: es muy complicado que se encuentren todas las tesis digitalizadas en los repositorios, debido a que varias todavía no han sido digitalizadas ni fueron distribuidas en formatos electrónicos, usualmente las más viejas.

[1] https://tesis.ipn.mx/discover

[2] https://www.repositoriodigital.ipn.mx/

Tabla 5-12. Producción de tesis por año.

Año	*Número de tesis*
2000	1
2001	0
2002	9
2003	17
2004	8
2005	8
2006	40
2007	46
2008	39
2009	54
2010	28
2011	35
2012	50
2013	45
2014	52
2015	26
2016	56
2017	76
2018	56
2019	39
2020	5

6. CONCLUSIONES

La importancia de un árbol de conocimiento para el sistema CLASSONTO es crucial, siendo éste la base para las categorías, por lo tanto, los fundamentos y su origen son esenciales, de esta manera si está basado en un estándar de una institución reconocida en el ámbito, brindará un consenso por parte de la comunidad; en este caso el árbol de conocimiento está basado en la ontología poli-jerárquica de ACM.

Las tesis analizadas, muy pocas contienen en su resumen las palabras clave (keywords), mucho menos están apegadas a un estándar de clasificación. Dando como resultado ambigüedad en algunos conceptos o una descripción imprecisa de la tesis. Por ejemplo, en el uso de la expresión "algoritmo de clasificación" para denotar una "técnica de aprendizaje no supervisado", siendo la terminología de "algoritmo de clasificación" para denotar una "técnica de aprendizaje supervisado". Por lo que uno esperaría encontrar en la tesis algoritmos de aprendizaje supervisado, cuando es completamente lo opuesto, siendo los algoritmos de agrupamiento (clustering) los utilizados.

Teniendo la posibilidad de modificar el árbol de conocimiento, se tiene una transparencia en el funcionamiento del sistema, siendo éste muy útil a la hora de lidiar con anglicismos y acrónimos (situaciones muy frecuentes en el ámbito de la computación), permitiendo que ambos representen clasificaciones, sin necesidad de usar etiquetas POS.

La utilidad del sistema no se limita a enumerar las clasificaciones, sino también a reconocer con que clasificaciones no cuenta la tesis. Permitiendo reconocer si aquellas clasificaciones que usualmente estén presentes de manera conjunta no lo están, e infiriendo lo que supone esto: posiblemente no uso una técnica complementaria, no uso una base de datos frecuente, etc. Por ejemplo: siendo la tesis un "Análisis de ruido acústico", y que las clasificaciones sean: "procesamiento digital de señales", "sistemas de información geográfica", "base de datos", pero que no contenga ninguna clasificación con respecto a "hardware"; se puede inferir que la tesis no hace un análisis del ruido desde la obtención de éste por medio de sensores.

El sistema tiene la posibilidad de clasificar temáticas que comparten vocabulario, por ejemplo:

Teniendo como título "Afectaciones a la salud por uso de comunicaciones móviles" la publicación a analizar. Y que trate de "afectaciones a la salud", que se relaciona con cáncer, que a su vez se relaciona con células. También tratará de "comunicaciones móviles", que a su vez se relaciona con celulares, y que estos a su vez tienen antenas

y fractales, que bien pudieron ser diseñadas por autómatas celulares. Siendo hasta tres clasificaciones que comparten vocabulario, y que se pueden clasificar en "bioinformática", "autómatas celulares" y "telefonía móvil". Y que son perfectamente identificables ya que enriqueciendo la terminología de la clasificación habrá más votaciones en las clasificaciones correctas.

Dado que el sistema está basado en el conteo de votos, el uso de propagación de votos sirve para "desambiguar" al momento de reconocer a que temáticas se refieren, sin importar si existen temáticas con vocabularios compartidos.

Por último, es importante categorizar la publicación científica, aprovechándolos como recursos digitales y brindando la posibilidad de su tratamiento automatizado, permitiendo búsquedas más eficientes ya sea para fines de evaluación del estado del arte, contrastar investigaciones, o evaluar el desarrollo de temas. A su vez de que existe la posibilidad de extender a diferentes dominios el uso del sistema CLASSONTO, empleando la ontología adecuada, por ejemplo, MarineTLO [1] (clasificación de especies marinas), THE GENE ONTOLOGY RESOURCE [2] (clasificación de sistemas biológicos desde nivel molecular hasta organismo), entre otras.

[1] https://projects.ics.forth.gr/isl/MarineTLO/

[2] http://geneontology.org/

7. TRABAJO A FUTURO

A continuación, se enlistan posibles entradas para un trabajo futuro:

- El trabajo se acotó a las tesis publicadas del CIC en el periodo 1996-2020, pero es posible tener el sistema CLASSONTO funcionando para que categorice conforme las tesis se vayan publicando, siendo posible montarlo en un servidor y que "sugiera" categorías de manera "automática".
- Los resultados obtenidos muestran que es posible clasificar las tesis teniendo como base un árbol de conocimiento; por tanto, es posible aplicar el clasificador a más instancias de publicaciones científicas como lo son: artículos, ensayos, etc.
- El sistema CLASSONTO al tener como base para la categorización un árbol de conocimiento, por sus características, éste permite una mejora continua, por lo que se podría ir expandiendo en términos, dando como resultado una mejor base para la categorización.
- También es posible aplicar el sistema CLASSONTO a otras instituciones que manejen el ámbito de la computación en sus publicaciones, permitiendo ampliar el árbol de conocimiento con terminología relacionada y accediendo a mayor material de publicaciones científicas para la evaluación del propio sistema CLASSONTO, ayudándose de la conceptualización compartida que representa una ontología.
- Este trabajo se desarrolló utilizando Python como lenguaje de programación; siendo este útil para el análisis de datos y lenguaje natural gracias a que existen librerías como NLTK (Natural Language Toolkit), Gensim, spaCy, FreeLing, etc. Sin embargo, se podría experimentar un aumento en rendimiento usando lenguajes compilados como lo es C.
- Muchos trabajos describen las técnicas usadas o propuestas con fórmulas matemáticas, por lo que se podría ejecutar un reconocimiento de dichas fórmulas para el reconocimiento de terminología; de igual manera se podría aplicar para las referencias bibliográficas.

APÉNDICE A

A.1 CLASIFICACIÓN ACM

La clasificación ACM consta de 13 categorías generales, a continuación, se enlistan y se describen algunas de las temáticas que engloban, con el fin de conocer mejor la clasificación.

1. General y referencia: Herramientas y técnicas de computación cruzada, estándar y guías.
2. Hardware: PCB, tecnologías emergentes, circuitos integrados, interfaces físicas.
3. Organización de sistemas computacionales: Arquitecturas paralelas y seriales, sistemas embebidos.
4. Redes: Tipos de red, arquitecturas de redes, algoritmos de redes, servicios en redes.
5. Software e Ingeniería de software: Organizaciones, compiladores, lenguajes de programación, repositorios.
6. Teoría de la computación: Aleatoriedad, lógica, probabilística, lenguajes formales, autómatas.
7. Matemáticas de la computación: Matemáticas discretas y continuas, teoría de la información.
8. Sistemas de información: Recuperación de la información, lenguajes de consulta, minería de datos.
9. Seguridad y privacidad: Criptografía, seguridad en redes, servicios.
10. Computación centrada al ser humano: Visualización, interacción, accesibilidad, colaboración.
11. Metodologías computacionales: Aprendizaje máquina, inteligencia artificial, modelado y simulación, computación paralela.
12. Computación aplicada: Comercio electrónico, artes y humanidades, educación, sociedad, medicina.
13. Temas de sociedad y profesionales: Políticas gubernamentales, censura, vigilancia.

A.2 GUÍA PARA CLASIFICAR EN ACM

ACM provee una serie de recomendaciones a la hora de clasificar una publicación en el sistema de clasificación computacional ACM.

- Ponderar la identificación de los nodos de más bajo nivel cuan sea posible. Entre más específico el concepto, mejor.
- Asignar ponderaciones a la clasificación, dependiendo de su relevancia en la publicación: Alta, Mediana, Baja.
- Obtener un mínimo de clasificaciones de relevancia Alta.
- Preferentemente usar clasificaciones de bajo nivel cuando sea aplicable, a menos que *ninguna* o *todas* apliquen, en dado caso se tendrá que usar la clasificación que las abarque (mayor nivel)
- Tener en cuenta que una publicación con clasificaciones generales sería una publicación que tiene temáticas muy generales y no específicas.
- Revisar la clasificación, si es correcta con el ámbito que manejan los nodos de alto nivel.

  Ejemplo: Si la publicación trata del "internet en términos de gobernanza":
  - La clasificación errónea sería:

    4. Redes

    4.8. Tipos de redes

    4.8.11. Internet público
  - Y la clasificación adecuada sería:

    13. Temas sociales y profesionales

    13.2. Política informática / tecnológica

    13.2.1. Propiedad intelectual

    13.2.1.6. Gobernanza de Internet / nombres de dominio

A.3 ÁRBOL DE CONOCIMIENTO BASADO EN LA CLASIFICACIÓN ACM

El árbol de conocimiento que es generado a partir de la ontología multi-jerárquica del ACM, ha sido enriquecido con 10592 términos (términos sin repetir en inglés o español). El sistema tiene la posibilidad de leer un archivo JSON o un archivo de texto con estructura similar, aunque con las ventajas que te brinda un TXT con viñetas como numeración de la jerarquía y la fácil visualización.

La estructura del archivo TXT es la siguiente:

Numeración Clasificación
{
"@es": ["*Términos en español*"],
"@en": ["*Términos en inglés*"]
}

Ejemplo:

2.3.4.1. Transistors
{
"@es": ["transistores", "transistor de efecto de campo"],
"@en": ["transistors", "FET", "field effect transistor"]
}

Este árbol de conocimiento se le aplica procesamiento de texto (eliminación de stopwords, lematización) para así tener términos consistentes en cada clasificación.

Para la visualización del árbol de conocimiento, se pone a disposición en formato de Microsoft Word (docx) en la siguiente liga: https://docs.google.com/document/d/1GLLFKw58JxweMiyTpZZpKmyYJqeiRpbe/edit?usp=sharing&ouid=115284057556858866741&rtpof=true&sd=true

De igual manera se puede poner en contacto a alberto.oscar96@gmail.com para solicitar una copia.

APÉNDICE B

B.1 EJECUCIÓN DE CLASSONTO

A continuación, se mencionan las scripts que conforman el sistema CLASSONTO y se describen sus funciones. El orden de lista tiene relación con el orden de ejecución.

1. "standard2metrics.py": Convierte el *Golden Standard* formato "json" a formato "csv".
2. "ontology_doc2json.py" (opcional):
    - Si el árbol de conocimiento es un archivo "txt", se realiza su lectura para guardarlo en formato "json" y posteriormente usarlo.
3. " ontology_json2dat.py ":
    - Convierte el archivo "json" a un diccionario para posteriormente guardarlos en un archivo "dat".
    - Obtiene estadísticas acerca del árbol de conocimiento.
    - Crea directorios con estructura similar al árbol, y en estos directorios muestra los términos procesados que votan por las categorías.
    - Valida los directorios creados.
4. "encode.py":
    - Analiza los componentes de interés (título, resumen, tesis completa), donde se establecen el valor que cada componente tendrá al final de la votación.
        - El análisis es a través del procesamiento de las tesis, y el posterior conteo de votos por medio de ventana.
        - El conteo de votos es de acuerdo a las coincidencias que se presenten en la tesis analizada y el árbol de conocimiento.
        - El tamaño de ventana está determinado por el máximo valor en la longitud de los términos.
    - Guarda el conteo de votos, los términos reconocidos y las palabras que no reconoció en archivos formato "dat"
5. "decode.py":
    - Convierte los archivos "dat" en archivos "csv" para una mejor visualización

6. "csv2metrics.py":
    - Realiza una normalización y poda (pruning) a las tablas de votaciones.
    - Realiza un mapeo de los valores, para obtener la importancia (Alta, Mediana, Baja) de las clasificaciones que corresponde con sus votos.
    - En dado caso que exista un Golden Standard (clasificaciones previamente realizadas) de la tesis analizada, se agrega como una columna a la tabla de votaciones, para su posterior comparación.
    - Fusiona toda la terminología desconocida de todas las tesis, indicando en que tesis aparecieron, y realiza una poda.
    - Guarda las modificaciones a las tablas de votaciones y a las palabras desconocidas en archivos "csv".
7. "validation.py":
    - Valida las tablas de votaciones con respecto al Golden Standard (con métricas propuestas y clasicas), y guarda el resultado en "csv".
8. "temporal_analysis.py":
    - Hace un análisis de las clasificaciones obtenidas de las tesis a lo largo del tiempo.
        - Genera graficas sobre las clasificaciones más repetidas en un determinado año.
        - Genera graficas sobres la evolución de clasificaciones seleccionadas.
        - Genera las respuestas a los cuestionamientos que se buscaban responder en el capítulo 1 (Visualización y Evaluación).
9. "inference_maker.py":
    - Realiza sugerencias de clasificaciones en la terminología relacionada desconocida, y guarda las sugerencias en el archivo que reúne todos los términos.

APÉNDICE C

C.1 MATRIZ DE CONFUSIÓN POR TESIS

La matriz confusión multi-etiqueta MCM tiene la cuenta de los valores Verdaderos Negativos $MCM_{0,0}$, Falsos negativos $MCM_{1,0}$, Verdaderos Positivos $MCM_{1,1}$, y Falsos Positivos $MCM_{0,1}$.

```
Confusion matrix for sample:  AcevedoMosquedaMariaElena_2006
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  AguilarGaliciaHonorato_2012
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  AlbortanteMoratoCecilia_2009
[[2101    9]
 [   2    1]]
Confusion matrix for sample:  AlcazarSilvaEliezer_2013
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  AlonsoCastroJoseAdriel_2017
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  AlvarezMendezJonatan_2014
[[2102    8]
 [   1    2]]
Confusion matrix for sample:  AvilaGamboaGuillermoIII_2017
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  BarceloAlonsoGrettel_2010
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  BarronFernandezRicardo_2006
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  BautistaBautistaPatricia_2009
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  CabreraRiveraLuis_2018
[[2102    4]
 [   1    6]]
Confusion matrix for sample:  CamachoEscotoJoseJaime_2017
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  CarreraTrejoJorgeVictor_2010
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  CastilloMontielErandi_2009
[[2101    8]
 [   2    2]]
```

```
Confusion matrix for sample:  CatalanSalgadoEdgarArmando_2007
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  CeronFigueroaSergio_2013
[[2103    4]
 [   0    6]]
Confusion matrix for sample:  CeronFigueroaSergio_2018
[[2102    7]
 [   1    3]]
Confusion matrix for sample:  CervantesRamirezAngelOmar_2012
[[2098    7]
 [   5    3]]
Confusion matrix for sample:  CoronaBermudezErendira_2020
[[2103    5]
 [   0    5]]
Confusion matrix for sample:  CortesAntonioPrometeo_2011
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  CruzSilvaJacoboEmanuel_2020
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  CuevasRasgadoAlmaDelia_2003
[[2101    7]
 [   2    3]]
Confusion matrix for sample:  DiazDiazJaime_2012
[[2101    7]
 [   2    3]]
Confusion matrix for sample:  DuchanoyMartinezCarlosAlberto_2012
[[2102    6]
 [   1    4]]
Confusion matrix for sample:  EstradaSanchezIvan_2010
[[2102    7]
 [   1    3]]
Confusion matrix for sample:  FarfanEstradaIsmael_2012
[[2101    8]
 [   2    2]]
Confusion matrix for sample:  FernandezCidHugoIvan_2019
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  FloresCortesAndres_2013
[[2102    5]
 [   1    5]]
Confusion matrix for sample:  FloresRoaAlberto_2010
[[2102    6]
 [   1    4]]
Confusion matrix for sample:  GarciaBlanquelEricka_2012
[[2098    8]
 [   5    2]]
Confusion matrix for sample:  GarciaBlanquelEricka_2017
[[2103    3]
 [   0    7]]
Confusion matrix for sample:  GarciaCortesMonica_2014
[[2102    6]
 [   1    4]]
Confusion matrix for sample:  GarciaLavanderosNorberto_2018
```

```
[[2102    8]
 [   1    2]]
Confusion matrix for sample:  GarroLiconBeatrizAurora_2012
[[2102    7]
 [   1    3]]
Confusion matrix for sample:  GermanSotoErnesto_2002
[[2103    5]
 [   0    5]]
Confusion matrix for sample:  GodinezFernandezEduardo_2009
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  GomezLuisAlexiaItzel_2019
[[2102    7]
 [   1    3]]
Confusion matrix for sample:  GonzalezGarciaAlainCesar_2007
[[2100    7]
 [   3    3]]
Confusion matrix for sample:  GutierrezSanchezAngelIvan_2015
[[2100    8]
 [   3    2]]
Confusion matrix for sample:  HernandezAcostaCindyGabriela_2018
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  HernandezCruzEstelaYadira_2020
[[2103    5]
 [   0    5]]
Confusion matrix for sample:  HernandezHernandezGerardo_2014
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  HernandezHernandezGerardo_2019
[[2102    7]
 [   1    3]]
Confusion matrix for sample:  HerreraLopezMariaGuadalupe_2018
[[2101    8]
 [   2    2]]
Confusion matrix for sample:  HuertaMorenoGabrielOmar_2009
[[2102    7]
 [   1    3]]
Confusion matrix for sample:  IbarraRomeroMartin_2014
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  IbarraVargasJoseJonathan_2009
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  JimenezArandaItzael_2018
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  JimenezGarciaGregorioArturo_2016
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  JuarezHipolitoJuanHumberto_2016
[[2102    7]
 [   1    3]]
Confusion matrix for sample:  JuarezMurilloCristianRemington_2012
[[2102    6]
```

```
 [   1    4]]
Confusion matrix for sample:  KolesnikovaOlga_2011
[[2102    6]
 [   1    4]]
Confusion matrix for sample:  LakeMoctezumaFranzLudwig_2019
[[2101    7]
 [   2    3]]
Confusion matrix for sample:  LavinVillaMoisesEduardo_2010
[[2100    7]
 [   3    3]]
Confusion matrix for sample:  LopezCardenasRodrigo_2008
[[2103    5]
 [   0    5]]
Confusion matrix for sample:  LopezPachecoMariaGuadalupe_2012
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  LopezVerasteguiGermanOswaldo_2014
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  LopezYañezItzama_2011
[[2101    7]
 [   2    3]]
Confusion matrix for sample:  MarquezMolinaMiguel_2013
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  MartinezCastilloJesusElohim_2012
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  MartinezCazaresEduardo_2009
[[2101    8]
 [   2    2]]
Confusion matrix for sample:  MartinezHernandezVictorManuel_2009
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  MartinezNavarroJoseAngel_2018
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  MartinezValleCarlosLeon_2006
[[2101    8]
 [   2    2]]
Confusion matrix for sample:  MataRiveraMiguelFelix_2009
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  MercadoCapistranPatricio_2015
[[2101    7]
 [   2    3]]
Confusion matrix for sample:  MontañoSanchezCesarEdgar_2016
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  MoralesFernandezAlecxis_2009
[[2102    6]
 [   1    4]]
Confusion matrix for sample:  NovoaCatañoJavier_2007
[[2102    6]
 [   1    4]]
```

```
Confusion matrix for sample:  NoyolaBautistaJoel_2014
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  OcampoPolitoRigoberto_2010
[[2102    6]
 [   1    4]]
Confusion matrix for sample:  OlveraOrtegaJorge_2003
[[2101    6]
 [   2    4]]
Confusion matrix for sample:  OropezaRodriguezJoseLuis_2005
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  PerezLeonJaimeAlfonso_2009
[[2101    8]
 [   2    2]]
Confusion matrix for sample:  PeñaAyalaAlejandro_2007
[[2102    7]
 [   1    3]]
Confusion matrix for sample:  QuinteroPerezGuillermo_2013
[[2103    6]
 [   0    4]]
Confusion matrix for sample:  QuinteroTellezRolando_2007
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  RenteriaAgualimpiaWalter_2009
[[2101    9]
 [   2    1]]
Confusion matrix for sample:  ReyesMartinezMiguelDaniel_2015
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  ReyesTorresJesusJavier_2017
[[2102    6]
 [   1    4]]
Confusion matrix for sample:  RiveraLozaGabriela_2003
[[2102    8]
 [   1    2]]
Confusion matrix for sample:  RodriguezMartinezMiguel_2011
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  RodriguezRomeroRaymundo_2020
[[2102    6]
 [   1    4]]
Confusion matrix for sample:  RojoRuizArturo_2008
[[2100    6]
 [   3    4]]
Confusion matrix for sample:  RomeroXimilJoseManuel_2002
[[2102    8]
 [   1    2]]
Confusion matrix for sample:  SanchezFragaRodolfo_2013
[[2102    8]
 [   1    2]]
Confusion matrix for sample:  SanchezFragaRodolfo_2016
[[2102    6]
 [   1    4]]
Confusion matrix for sample:  SuarezOropezaMiguel_2012
```

```
[[2102    7]
 [   1    3]]
Confusion matrix for sample:  TorresCruzNoe_2019
[[2102    6]
 [   1    4]]
Confusion matrix for sample:  TrejoSotoGloriaIrene_2002
[[2102    6]
 [   1    4]]
Confusion matrix for sample:  UriarteArciaAbrilValeria_2012
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  UriarteArciaAbrilValeria_2016
[[2101    9]
 [   2    1]]
Confusion matrix for sample:  ValleChavezAbel_2012
[[2102    7]
 [   1    3]]
Confusion matrix for sample:  VanegasSanchezTonatiuhDaniel_2018
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  VarelaGarciaFrancisco_2002
[[2101    8]
 [   2    2]]
Confusion matrix for sample:  VazquezOropezaJonathan_2016
[[2101    7]
 [   2    3]]
Confusion matrix for sample:  VelazquezAlcantaraRodrigoAlejandro_2017
[[2101    5]
 [   2    5]]
Confusion matrix for sample:  VerasteguiBarrancoKarina_2007
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  ZarateEscobedoRicardo_2018
[[2103    7]
 [   0    3]]
Confusion matrix for sample:  ZhilaAlisa_2014
[[2103    6]
 [   0    4]]
```

APÉNDICE D

D.1 MATRIZ DE CONFUSIÓN POR CLASIFICACIÓN

Las clasificaciones mostradas, son aquellas que están presentes en el *Golden Standard* por lo que no constituyen la totalidad de clasificaciones en el árbol de conocimiento. La matriz confusión multi-etiqueta MCM tiene la cuenta de los valores Verdaderos Negativos $MCM_{0,0}$, Falsos negativos $MCM_{1,0}$, Verdaderos Positivos $MCM_{1,1}$, y Falsos Positivos $MCM_{0,1}$.

```
Confusion Matrix for class:  2.
[[90  9]
 [ 0  1]]
Confusion Matrix for class:  2.2.
[[93  6]
 [ 0  1]]
Confusion Matrix for class:  2.2.1.
[[94  3]
 [ 0  3]]
Confusion Matrix for class:  2.2.1.1.
[[88  5]
 [ 1  6]]
Confusion Matrix for class:  2.2.2.
[[90  3]
 [ 0  7]]
Confusion Matrix for class:  2.2.4.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  2.2.8.
[[97  0]
 [ 2  1]]
Confusion Matrix for class:  2.2.9.
[[98  0]
 [ 1  1]]
Confusion Matrix for class:  2.2.10.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  2.2.15.
[[98  0]
 [ 0  2]]
Confusion Matrix for class:  2.3.
[[93  6]
 [ 0  1]]
Confusion Matrix for class:  2.3.4.1.
[[97  2]
 [ 0  1]]
Confusion Matrix for class:  2.3.5.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  2.3.6.
[[96  1]
 [ 0  3]]
Confusion Matrix for class:  2.3.6.3.
[[95  2]
 [ 0  3]]
Confusion Matrix for class:  2.3.6.5.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  2.6.1.2.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  2.6.2.
[[98  1]
 [ 0  1]]
Confusion Matrix for class:  2.7.1.5.
[[96  0]
 [ 4  0]]
Confusion Matrix for class:  2.7.3.1.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  2.8.9.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  2.9.3.5.
[[96  1]
 [ 1  2]]
Confusion Matrix for class:  2.10.4.1.
[[98  1]
 [ 0  1]]
Confusion    Matrix    for    class:
2.10.10.4.
[[93  0]
 [ 0  7]]
Confusion Matrix for class:  3.1.2.
[[98  0]
 [ 0  2]]
Confusion Matrix for class:  3.1.3.
[[95  3]
 [ 0  2]]
Confusion Matrix for class:  3.1.4.1.
```

```
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  3.1.4.2.
[[96  3]
 [ 0  1]]
Confusion Matrix for class:  3.1.4.8.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  3.2.2.
[[94  5]
 [ 0  1]]
Confusion Matrix for class:  3.2.2.1.
[[97  1]
 [ 0  2]]
Confusion Matrix for class:  3.2.2.2.
[[93  2]
 [ 1  4]]
Confusion Matrix for class:  3.2.3.
[[90  9]
 [ 0  1]]
Confusion Matrix for class:  3.2.4.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  4.2.
[[87 10]
 [ 0  3]]
Confusion Matrix for class:  4.2.3.
[[98  1]
 [ 0  1]]
Confusion Matrix for class:  4.2.4.1.
[[96  2]
 [ 0  2]]
Confusion Matrix for class:  4.2.6.
[[98  1]
 [ 0  1]]
Confusion Matrix for class:  4.4.1.
[[98  1]
 [ 0  1]]
Confusion Matrix for class:  4.4.1.3.
[[97  2]
 [ 0  1]]
Confusion Matrix for class:  4.6.1.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  4.6.1.1.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  4.6.1.2.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  4.6.5.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  4.8.6.1.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  4.8.8.1.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  4.8.9.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  4.8.10.
[[96  2]
 [ 0  2]]
Confusion Matrix for class:  4.8.10.1.
[[97  1]
 [ 0  2]]
Confusion Matrix for class:
5.1.1.2.3.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  5.1.1.3.
[[94  5]
 [ 0  1]]
Confusion Matrix for class:
5.1.1.3.2.
[[98  0]
 [ 1  1]]
Confusion Matrix for class:
5.1.2.5.2.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:
5.1.3.2.2.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  5.2.1.2.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  5.2.3.10.
[[97  1]
 [ 0  2]]
Confusion Matrix for class:
5.2.4.1.1.
[[97  2]
 [ 0  1]]
Confusion Matrix for class:  5.2.4.8.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  5.2.8.
[[97  2]
 [ 0  1]]
Confusion Matrix for class:  5.3.2.1.
[[98  0]
 [ 0  2]]
Confusion Matrix for class:  5.3.3.
[[97  2]
 [ 0  1]]
Confusion Matrix for class:  5.3.4.
```

```
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  6.1.6.1.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  6.3.
[[98  0]
 [ 0  2]]
Confusion Matrix for class:  6.5.2.1.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  6.5.2.3.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  6.5.3.
[[94  4]
 [ 1  1]]
Confusion    Matrix    for    class:
6.5.3.1.2.2.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  6.5.3.2.
[[99  0]
 [ 1  0]]
Confusion    Matrix    for    class:
6.5.3.2.8.
[[98  1]
 [ 0  1]]
Confusion    Matrix    for    class:
6.5.3.3.3.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  6.6.4.
[[98  1]
 [ 0  1]]
Confusion    Matrix    for    class:
6.7.1.4.1.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  6.7.1.6.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  6.7.3.6.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  6.7.3.9.
[[98  0]
 [ 1  1]]
Confusion Matrix for class:  6.8.
[[84  5]
 [ 2  9]]
Confusion Matrix for class:  7.1.2.9.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  7.2.
[[97  0]
 [ 2  1]]
Confusion Matrix for class:  7.2.3.5.
[[98  1]
 [ 1  0]]
Confusion Matrix for class:  7.2.3.7.
[[96  1]
 [ 0  3]]
Confusion Matrix for class:  7.2.5.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  7.2.6.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  7.4.1.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  7.5.1.1.
[[97  1]
 [ 0  2]]
Confusion Matrix for class:  7.5.1.8.
[[99  0]
 [ 0  1]]
Confusion    Matrix    for    class:
7.5.2.1.2.2.
[[99  0]
 [ 0  1]]
Confusion    Matrix    for    class:
7.5.2.2.8.
[[99  0]
 [ 0  1]]
Confusion    Matrix    for    class:
7.5.2.3.3.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  7.5.7.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  8.1.1.
[[98  1]
 [ 0  1]]
Confusion    Matrix    for    class:
8.1.1.5.2.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  8.1.3.2.
[[87  9]
 [ 0  4]]
Confusion Matrix for class:  8.1.3.5.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  8.1.4.2.
[[99  0]
 [ 1  0]]
```

```
Confusion Matrix for class: 8.1.4.2.1.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  8.1.6.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  8.1.6.2.
[[98  0]
 [ 0  2]]
Confusion Matrix for class:  8.1.6.7.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  8.3.
[[92  6]
 [ 1  1]]
Confusion Matrix for class:  8.3.2.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  8.3.3.
[[88 11]
 [ 0  1]]
Confusion Matrix for class:  8.3.3.1.
[[96  3]
 [ 0  1]]
Confusion Matrix for class:  8.3.3.2.
[[87  0]
 [ 1 12]]
Confusion Matrix for class:  8.3.3.4.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  8.3.7.2.
[[98  1]
 [ 0  1]]
Confusion Matrix for class:  8.3.8.4.
[[93  6]
 [ 0  1]]
Confusion Matrix for class:  8.3.8.6.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  8.3.9.
[[98  0]
 [ 0  2]]
Confusion Matrix for class:  8.4.3.2.
[[98  0]
 [ 0  2]]
Confusion Matrix for class:  8.4.3.4.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  8.4.6.
[[97  2]
 [ 0  1]]
Confusion Matrix for class:  8.4.7.1.
[[92  3]
 [ 0  5]]
Confusion Matrix for class: 8.4.7.1.1.
[[96  3]
 [ 0  1]]
Confusion Matrix for class:  8.4.7.2.
[[99  0]
 [ 0  1]]
Confusion Matrix for class: 8.4.7.2.1.
[[97  1]
 [ 0  2]]
Confusion Matrix for class:  8.5.
[[94  4]
 [ 0  2]]
Confusion Matrix for class:  8.5.1.6.
[[89  4]
 [ 0  7]]
Confusion Matrix for class:  8.5.1.7.
[[95  3]
 [ 1  1]]
Confusion Matrix for class:  8.5.4.2.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  8.5.4.4.
[[95  2]
 [ 0  3]]
Confusion Matrix for class:  8.5.5.
[[90  9]
 [ 0  1]]
Confusion Matrix for class:  8.5.5.2.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  8.5.5.4.
[[97  1]
 [ 0  2]]
Confusion Matrix for class:  8.5.5.5.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  8.5.5.8.
[[69 19]
 [ 0 12]]
Confusion Matrix for class: 8.5.8.1.1.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  9.1.
[[97  0]
 [ 0  3]]
Confusion Matrix for class:  9.1.2.
[[97  2]
 [ 0  1]]
Confusion Matrix for class:  9.3.1.
[[98  0]
 [ 0  2]]
Confusion Matrix for class:  9.4.1.
```

```
[[97  0]
 [ 0  3]]
Confusion Matrix for class:  9.7.1.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  10.1.2.5.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  10.1.3.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  10.1.3.4.
[[97  2]
 [ 0  1]]
Confusion Matrix for class:  10.1.3.6.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  10.2.1.1.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  10.4.3.
[[95  4]
 [ 0  1]]
Confusion Matrix for class:  10.4.3.1.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  10.4.3.3.
[[96  3]
 [ 1  0]]
Confusion Matrix for class:  10.4.3.4.
[[98  1]
 [ 0  1]]
Confusion Matrix for class:  10.5.
[[98  1]
 [ 0  1]]
Confusion Matrix for class:  10.5.2.3.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  10.5.6.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  11.2.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  11.2.1.3.
[[98  0]
 [ 2  0]]
Confusion Matrix for class:  11.2.2.
[[97  1]
 [ 1  1]]
Confusion Matrix for class:  11.3.
[[80 19]
 [ 0 11]]
Confusion Matrix for class:  11.3.1.3.
[[97  2]
 [ 1  0]]
Confusion Matrix for class:  11.3.1.4.
[[98  0]
 [ 0  2]]
Confusion Matrix for class:  11.3.1.5.
[[96  1]
 [ 0  3]]
Confusion Matrix for class:  11.3.1.6.
[[95  2]
 [ 1  2]]
Confusion Matrix for class:  11.3.1.7.
[[97  2]
 [ 0  1]]
Confusion Matrix for class:  11.3.1.8.
[[98  0]
 [ 1  1]]
Confusion Matrix for class:  11.3.2.
[[94  5]
 [ 0  1]]
Confusion Matrix for class:  11.3.2.2.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  11.3.2.5.
[[90  5]
 [ 0  5]]
Confusion Matrix for class:  11.3.2.9.
[[96  3]
 [ 0  1]]
Confusion Matrix for class:  11.3.4.1.
[[97  2]
 [ 1  0]]
Confusion Matrix for class:  11.3.5.2.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  11.3.5.3.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  11.3.8.
[[96  1]
 [ 0  3]]
Confusion Matrix for class:  11.3.8.1.
[[97  2]
 [ 0  1]]
Confusion    Matrix    for    class:
11.3.8.1.5.
[[99  0]
 [ 0  1]]
Confusion    Matrix    for    class:
11.3.8.1.7.
[[97  0]
 [ 3  0]]
```

```
Confusion Matrix for class: 11.3.8.2.6.
[[99  0]
 [ 0  1]]
Confusion Matrix for class: 11.3.8.3.1.
[[99  0]
 [ 1  0]]
Confusion Matrix for class: 11.3.8.4.2.
[[97  0]
 [ 1  2]]
Confusion Matrix for class: 11.3.8.4.9.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  11.4.
[[86 10]
 [ 0  4]]
Confusion Matrix for class:  11.4.1.
[[89  5]
 [ 1  5]]
Confusion Matrix for class:  11.4.1.1.
[[80  8]
 [ 0 12]]
Confusion Matrix for class: 11.4.1.1.1.
[[99  0]
 [ 1  0]]
Confusion Matrix for class: 11.4.1.1.3.
[[94  3]
 [ 1  2]]
Confusion Matrix for class:  11.4.1.2.
[[92  5]
 [ 0  3]]
Confusion Matrix for class:  11.4.3.
[[88 11]
 [ 0  1]]
Confusion Matrix for class: 11.4.3.2.1.
[[94  3]
 [ 0  3]]
Confusion Matrix for class:  11.4.3.3.
[[81  6]
 [ 0 13]]
Confusion Matrix for class: 11.4.3.4.1.
[[99  0]
 [ 0  1]]
Confusion Matrix for class: 11.4.3.5.5.
[[99  0]
 [ 0  1]]
Confusion Matrix for class: 11.4.3.5.6.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  11.4.3.6.
[[95  4]
 [ 0  1]]
Confusion Matrix for class: 11.4.3.6.1.
[[95  3]
 [ 1  1]]
Confusion Matrix for class:  11.4.3.8.
[[99  0]
 [ 1  0]]
Confusion Matrix for class: 11.4.3.11.
[[99  0]
 [ 1  0]]
Confusion Matrix for class: 11.4.3.14.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  11.4.4.
[[98  1]
 [ 0  1]]
Confusion Matrix for class:  11.4.4.3.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  11.5.
[[88  6]
 [ 1  5]]
Confusion Matrix for class:  11.5.1.
[[96  1]
 [ 2  1]]
Confusion Matrix for class:  11.5.3.
[[93  4]
 [ 1  2]]
Confusion Matrix for class:  11.5.3.5.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  11.5.3.6.
[[90  2]
 [ 0  8]]
Confusion Matrix for class: 11.5.3.13.
[[99  0]
 [ 1  0]]
Confusion Matrix for class:  11.6.3.2.
[[98  1]
 [ 0  1]]
Confusion Matrix for class:  12.2.8.
[[88 10]
 [ 0  2]]
Confusion Matrix for class: 12.2.15.2.
```

```
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  12.3.1.1.
[[96  2]
 [ 0  2]]
Confusion Matrix for class:  12.3.5.
[[96  2]
 [ 0  2]]
Confusion Matrix for class:  12.3.5.1.
[[97  2]
 [ 0  1]]
Confusion Matrix for class:  12.3.9.
[[95  4]
 [ 0  1]]
Confusion Matrix for class:  12.4.
[[94  3]
 [ 0  3]]
Confusion Matrix for class:  12.4.6.
[[97  1]
 [ 2  0]]
Confusion Matrix for class:  12.4.9.2.
[[98  1]
 [ 0  1]]
Confusion Matrix for class:  12.5.4.
[[98  1]
 [ 0  1]]
Confusion Matrix for class:  12.8.2.
[[98  0]
 [ 0  2]]
Confusion Matrix for class:  12.9.5.
[[94  3]
 [ 0  3]]
Confusion Matrix for class:  12.10.
[[99  0]
 [ 1  0]]
Confusion    Matrix    for    class:
13.1.2.1.5.
[[98  1]
 [ 0  1]]
Confusion Matrix for class:  13.1.2.3.
[[99  0]
 [ 0  1]]
Confusion Matrix for class:  13.2.4.
[[98  0]
 [ 0  2]]
Confusion Matrix for class:  13.2.7.4.
[[97  2]
 [ 0  1]]
```