BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA
FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

APLICACIÓN DE MINERIA DE DATOS A ESTUDIOS HISTÓRICOS PROSOPOGRÁFICOS

T E S I S

PARA OBTENER EL TÍTULO DE
INGENIERO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA
LUIS ALBERTO MARTÍNEZ ÁLVAREZ

ASESOR
Dr. Ivo Humberto Pineda Torres
Facultad De Ciencias de la Computación

COASESORA
Dra. Alicia Tecuanhuey Sandoval
Instituto de C.S y Humanidades “Alfonso Velez Pliego”

2009

VIEP

RESUMEN

La prosopografía es una técnica utilizada dentro de la investigación histórica cuyo objetivo es el análisis masivo de biografías de sujetos pertenecientes a un determinado grupo social, generalmente élites sociales o políticas, para detectar subconjuntos conformados a partir de variables socioeconómicas, socio profesionales, demográficas y políticas con el objeto de construir modelos y tendencias dentro de un colectivo.

Dada la gran cantidad de información que puede ser recopilada por los historiadores, esta se organiza en bases de datos. Se establece un análisis y aplicación de algoritmos de minería de datos que faciliten la obtención de resultados y tendencias dentro de la investigación histórica.

ÍNDICE

1. Definición del proyecto

1.1 Introducción

La fuerte presencia que las ciencias de la computación y las tecnologías de la información han ido tomando en muchas áreas del desarrollo humano, obligan a crear vínculos interdisciplinarios fuertes que coadyuven en la implementación de nuevas formas y metodologías de trabajo para la resolución de problemas.

Las ciencias de la computación actualmente tienen impacto en un sinnúmero de disciplinas en algunos casos tan complejas como dispares, las tecnologías de información se encuentran presentes en el sector productivo, educativo, científico, público y comercial entre muchos otros. En el área de las ciencias sociales y humanidades también se ha venido desarrollando cierta integración, aunque muy lentamente algunas veces con casos de éxito muy importantes y en algunos otros con resultados pocos satisfactorios.

La disciplina histórica en su interés por afinar sus métodos de análisis e interpretación de las huellas del pasado, para estudiar los múltiples aspectos del devenir de las sociedades humanas ha requerido de metodologías más sofisticadas y precisas que le permitan manejar información voluminosa susceptible de análisis cuantitativo.

La tarea del historiador que hoy día consiste en la comprensión de fenómenos humanos del pasado, implica una serie de elecciones, decisiones y apreciaciones intelectuales sobre la materia prima: los documentos. El tratamiento simplificado e ingenuo en estos materiales por largo tiempo le dieron a esta disciplina un carácter conjetural. La disciplina que une el estudio de los muertos a los vivos, y que continúa atada a lo que las evidencias puedan sustentar se ha desarrollado, por lo que tales procedimientos son más complejos. A la diversidad de métodos cualitativos para el examen de los documentos que surgieron en el siglo XV con los análisis filosóficos, se le ha unido el análisis cuantitativo en un compromiso más estrecho con la veracidad.

La posibilidad de contar, cifrar, calcular, compensa la incertidumbre que el historiador siempre enfrenta ante las fuentes de información. Los historiadores profesionales saben que los documentos no son evidencias fieles de su tiempo. Fueron elaborados con una intencionalidad y pensados para un público. Su conservación en archivo, también supuso la intervención de una subjetividad que no está despejada de parcialidad.

La representación cuantitativa de fenómenos históricos, impone a esta disciplina, un mayor rigor y eficacia a la metodología cualitativa, aunque no la sustituye por cuanto no toda información puede ser reductible a series homogéneas y comparables.

De esta forma el historiador no solo comprende la revisión exhaustiva de un sinfín de documentos históricos (testamentos, actas, libros, periódicos, etc.) para obtener datos específicos que formen una serie cronológica que le ahorra tiempo para su clasificación, ordenamiento y procesamiento. La cuantificación le obliga a decidir los conceptos implícitos en las clasificaciones así como en sus hipótesis.

La implementación de bases de datos es actualmente la solución más simple y eficaz para las necesidades de clasificación y acceso a la información y si a ello aumentamos el subsecuente análisis de los datos para la extracción de un conocimiento o tendencia a partir de ellos, las soluciones y resultados obtenidos pueden ser muy fructíferos.

Nos encontramos pues ante una oportunidad más de integración de las ciencias de la computación, en este caso con las ciencias sociales, y la historia en particular, para el auxilio en las tareas de investigación histórica. ¿De que manera?, implementando un conjunto de soluciones y algoritmos que permitan extraer conocimiento a partir de la integración de la información resultante del trabajo de historiadores en sistemas de información histórica.

El presente trabajo con el objetivo de presentar la línea lógica de la integración de la problemática histórica y computacional está organizado de la siguiente manera.

Capítulo 1: Presenta una introducción conceptual del tema a tratar, sus antecedentes y situación actual, así como los objetivos a ser cubiertos en el desarrollo del mismo.

Capítulo 2: Introduce y explica de manera clara y sintetizada los procesos de investigación histórica, y la definición de prosopografía como metodología específica de investigación.

Capítulo 3: Describe, y desarrolla los sistemas de información Histórica como parte inicial del proceso de integración.

Capítulo 4: Presenta a manera de marco teórico la definición de minería de datos, sus clasificaciones, algoritmos e implementaciones.

Capítulo 5: Desarrollo del caso de estudio en particular, el proyecto de Autonomía, el desarrollo e implementación de bases de datos y el diseño del SIH.

Capitulo 6: Analiza, selecciona e implementa los algoritmos de minería de datos propios para el proyecto de investigación.

Capítulo 7: Muestra los resultados obtenidos, y su análisis

Capítulo 8: Presenta las conclusiones y propuestas para el trabajo futuro

Se incluyen un par de apéndices el primero enlista y describe brevemente el conjunto de programas, utilidades y software utilizado para el desarrollo del proyecto, el

segundo muestra brevemente la interface del Sistema de Información Histórica implementado para el caso de estudio.

1.2 Descripción del Proyecto

1.2.1 Descripción del proyecto origen

El presente trabajo surge a partir de la necesidad de clasificar e integrar la información resultante de un arduo trabajo de investigación histórica emprendido por la Dra. Alicia Tecuanhuey Sandoval titulado “Las Bases Sociales del Autonomismo Poblano 1808-1835” cuyo interés particular es el de analizar las bases sociales del llamado autonomismo que pretende analizar las tendencias políticas que buscaron ejercer la autonomía en la jurisdicción de la provincia y posteriormente en el Estado de Puebla.

Para efectos de esta investigación el estudio comienza a partir de la selección de un grupo de personas en este caso definido por la función política que cumplen, es decir como representantes populares en diferentes niveles de gobierno en un Estado de provincia en particular Puebla. Este proyecto hace uso de la prosopografía para el análisis del grupo seleccionado por lo que la información obtenida, es valiosa y se convierte en una amplia fuente de conocimiento.

La historiadora conocedora de las capacidades y alcances de la tecnología actual, planteó la posibilidad de integrar dicha información en una base de datos por lo que con ayuda de sus colaboradores elaboró un primer intento, vaciando sus datos en una tabla de Microsoft Access®. La magnitud de estos datos y las posibilidades de análisis que sobre ellos podría realizarse abrió la posibilidad de plantearse un proyecto más ambicioso en cuanto a la clasificación y análisis de la información, que coadyuvara en la obtención de resultados y estadísticas que den más peso a la hipótesis planteada por la investigadora en el proyecto origen.

Las técnicas de minería de datos se plantean, entonces como una posibilidad para la obtención más eficaz y concreta de conocimiento y estadísticas sobre la información obtenida. Se plantea la aplicación de minería de datos a estudios históricos prosopográficos, no sólo como un método para lo obtención de gráficos y estadísticas, sino como el estudio y aplicación de los algoritmos de minería de datos, y el planteamiento del diseño y uso de sistemas de información histórica como herramienta importante dentro del proceso de investigación histórica en aquellas investigaciones que su naturaleza lo permita.

1.2.2 Antecedentes

A finales del siglo XIX y principios del XX, la cuantificación llegó a la historia moderna a través de los modelos positivistas de aquellos años, pero no fue sino hasta la década de los 60's y sobre todo en la de los 70´s, cuando el desarrollo de la informática hizo posible el tratamiento de grandes volúmenes de información, abriendo la posibilidad a historiadores a realizar análisis en fuentes de datos seriadas que puedan ser sometidas más fácilmente a una normalización. (Fernández. F. 2000)

De esta manera se encontraron los intereses de varias escuelas de historiadores, por un lado, los que trataban de construir una historia total bajo la línea francesa de Annales utilizando la cuantificación y la historia serial como método para la extracción de conclusiones, que coincide con la corriente de New Economic History. Por otro aquellos que plantean la introducción de los métodos de las ciencias sociales en la investigación histórica, además de aquellos que plantean el uso de la computadora para el análisis lingüístico. Esto generaba polémica dado el planteamiento de algunas tesis que se trataban de demostrar mediante la aplicación de los nuevos recursos de cálculo.

Comenzaron a surgir los primeros manuales y cursos que introducían a los historiadores en el mundo de la informática y el manejo de computadoras, la computadora ya como una nueva herramienta de trabajo disponible para el historiador se consolidaba como tal, haciendo más fácil la tabulación, ordenación y recuento de grandes cantidades de datos históricos, imposible con los métodos anteriores, las primeras publicaciones periódicas especializadas en la difusión de la metodología y de los resultados obtenidos al emplear informática en el tratamiento y análisis de los datos históricos aparecen.

A Principios de los 80's con el trabajo de historiadores en algunas áreas como la historia económica, la demografía o el estudio de estructuras sociales que podían ser objeto de cuantificación se desarrollaron sistemas de cálculo automática que fueron pioneros para otras iniciativas. Aunque gran parte del trabajo con computadora por parte de historiadores se vio limitado por la complejidad de los medios de almacenamiento (tarjetas perforadas con capacidad de una o dos líneas de texto) y los complejos sistemas de cálculo que hacían muy difícil el manejo de la información, obligando a historiadores a contratar especialistas para introducir información en miles de tarjetas, desalentando a muchos de ellos en incluir la informática como herramienta viable para auxilio de su investigación.

A finales de esta década se pusieron en marcha asociaciones internacionales como la Internacional Association For History and Computing que coordinaran esfuerzos,

organizaran eventos y definieran soluciones metodológicas que tuvieran que ver con la aplicación de la informática para resolver problemas comunes en el ámbito de investigación histórica. Es también cuando, con la aparición del concepto de computadora personal, y el acercamiento de una diversidad de herramientas ofimáticas, como procesadores de palabras, gestores de base de datos, hojas de cálculo, paquetes de análisis estadístico, gestores de archivos y demás herramientas, hubo mayor acercamiento por parte de los historiadores con la informática.

Este nuevo impulso dio principio a la integración sistemática de material de trabajo en proyectos de mayor envergadura; se trabajó en otros casos como el análisis textual y la explotación mediante computadora de determinados tipos de documentos; la aplicación de la informática también permitió el análisis de datos sociales para construcciones historiográficas como es el caso de la biografía colectiva o prosopografía. Como es el caso de trabajos exitosos como:

Calvo Cuenca, Antonio; Jiménez Ruiz, Alfonso; Serrano Tenlíado, María Araceli: «Bases de datos relacionales para el análisis e interpretación de fuentes notariales en Historia Moderna: ventajas, limitaciones y perspectivas de futuro» LHEUNF, 2000, »Bernardo Ares, José Manuel de: «Informatización del Trabajo Científico Bibliográfico y Documental (INTRACIBI-DO)» LIIEUNE, 2000, " Ostolaza Elizondo, Maria Isabel: «Fuentes de información del Consejo de Estado. Base de datos sobre el Consejo de Estado y Navarra en los siglos xví-xvíí», LHEUNF, 2000, ' Montiel Torres, María Francisca; Villas Tinoco, Siro: «Propuesta para un modelo de análisis automatizado de redes sociales de interés en la Edad Moderna», LHEUNF, 2000. p. 187. SánchezBalmaseda, María Isabel: «Análisis de Redes Sociales: una herramienta en manos de los historiadores », LIIEUNF, 2000 Zofio Llorente. Juan Carlos: «Aplicación de Bases de Datos relacionales en la investigación histórica: familia y oficio en la Edad Moderna», LI-IFUNF, 2000, Lema Pueyo, José Ángel, y Munito Loinaz, José Antonio: «Nuevos documentos y nuevotratamiento de las fuentes para el estudio de la lucha de bandos» en Díaz de Durana de Urbina, José Ramón: La lucha de bandos en el País Vasto: de los parientes mayores a la hidalguía universal, Guipózcoa, de los bandos a la provincia (siglos XIV a XVI), Bilbao: Servicio editorial. Universidad del País Vasco/Euskal Herriko Unibertsitatea, 1998. Pagarolas Sabaté, Laureá: «Laplicació de la informática sobre els llibres notarials a lArxiu de Protocols de Barcelona»Lligalí, 4, 1991.

La llegada y evolución de la tecnología en cuanto a almacenamiento, poder de procesamiento, interconectividad y accesibilidad, supuso un cambio sustancial en el empleo de la informática. Mientras la aplicación de la informática suponía grandes gastos de dinero y tiempo, se invertía esfuerzo en una minuciosa programación de las tareas a que iban a ser sometidas las informaciones recogidas. Esto requería que en el momento de la publicación de los resultados se tuviera que explicar fuertemente el método informático también, Con la evolución y la mayor facilidad en el uso de

aplicaciones informáticas, los métodos informáticos fueron pasando a segundo término y el uso de las herramientas informáticas se hizo muy cotidiano a tal punto que su uso es intrínseco en algunas operaciones, se dejó a un lado la profundización sobre herramientas y técnicas informáticas, a cambio del uso masivo de herramientas estadísticas o comerciales.

Actualmente la aplicación de la informática y la historia se encuentra enfocada principalmente en la catalogación y clasificación de material bibliográfico, en el manejo de fuentes y en la archivística como el CISOC o el CEDIC.

1.2.3 Objetivos

El objetivo principal que persigue este trabajo es establecer, seleccionar y aplicar un conjunto de algoritmos de minería de datos que ayuden en el proceso de investigación histórica denominado prosopografía, a partir de la construcción de un sistema de información histórica y el análisis de los datos perfectamente estructurados dentro de él. Lo anterior da origen a un conjunto de objetivos específicos:

- ✓ Analizar, estudiar y aprender a aplicar los diferentes algoritmos y herramientas de minería de datos.

- ✓ Conocer el funcionamiento y antecedentes de los Sistemas de Información Histórica, para el desarrollo de uno propio en el caso de estudio particular

- ✓ Aprender los pasos a seguir dentro de un proceso de investigación histórica en particular del análisis de biografías colectivas y prosopografía.

- ✓ Desarrollar un análisis de la información recopilada por los historiadores, clasificarla y generar la estructura de una base de datos relacional, sólida y balanceada.

- ✓ Proporcionar al historiador un conjunto de herramientas para la captura y visualización de los datos.

- ✓ Generar un conjunto de estadísticas, gráficos y datos extraídos a partir de la aplicación de algoritmos de minería de datos.

2. Prosopografía

2.1 Introducción

La disciplina histórica como una rama del conocimiento, adquiere una importancia fundamental en el desarrollo humano y es base fundamental en la cultura de todo profesional, independientemente de su especialidad, además es parte importante de la formación de los distintos tipos de ciudadanos de cada país. Todo miembro perteneciente a una comunidad social encuentra en las distintas versiones del pasado , los conocimientos necesarios comprender su presente, para reflexionar sobre la variedad de respuestas humanas a retos del pasado similares al presente y para imaginar el futuro.

La disciplina histórica que hoy día es profesional ha evolucionado tanto en definir la función social que desempeña en las comunidades humanas como en sus procedimientos. Por ello los historiadores a lo largo del tiempo han venido transformando los alcances de su labor. En principio se profundizaba más en el contenido y objetivo literario posteriormente se llegó a pensar en relación con la verdad y finalmente se ha preocupado por una mayor vigilancia epistemológica. De esta forma la disciplina histórica se diferenció de otras formas del recuerdo del pasado, tales como la leyenda, mitos, relatos épicos, etc.

2.2 El proceso de investigación histórica

La investigación histórica tiene como etapas principales

1. El planteamiento y enunciación del problema

   El principio de un proyecto de investigación comienza cuando se pretende analizar y comprender un hecho histórico o experiencia del pasado, el investigador entonces aún con una noción vaga de los alcances de la interrogante intenta aislar los elementos fundamentales que suscitan la inquietud y verifica que existan los métodos de indagación y las fuentes de investigación, hasta plantear un enunciado simple y claro.

2. La recolección del material informativo

   El investigador explora todas las fuentes y testimonios a su alcance, y selecciona aquellos que se relacionan con el problema. El universo de fuentes y documentos es muy amplio para esta investigación concreta como Actas legislativas, judiciales, ejecutivas, cédulas, actas de nacimiento, testamentos, datos conservados por las iglesias, actas de bautizo, de matrimonio, actas notariales. Entre muchos otros.

3. La crítica de lo los datos acumulados

Se debe comprobar la autenticidad, valides y trascendencia de las fuentes por lo que se plantean diferentes cuestionamientos como la intención para elaborar el documento, quien fue el autor del documento, en que contexto lo produjo, si es el original o una copia, cuando, donde y porqué fue producido.

4. La formulación de hipótesis para explicar los diversos hechos o condiciones

   Con base en el conocimiento del contexto, la bibliografía especializada y un primer sondeo sobre las fuentes se proponen diferentes hipótesis que expliquen sucesos y condiciones, buscan conexiones ocultas, pautas, y se procura explicar las interrelaciones de estructura en los fenómenos, después de formular estas hipótesis se buscan las pruebas que las confirmen o refuten.

5. La interpretación de los descubrimientos y redacción del informe

   Una vez recolectadas todas las fuentes, analizadas y que se han completado los descubrimientos se procede a la redacción de los informes que expongan el desarrollo del proyecto, en tal exposición se incluye el enunciado inicial del proyecto, reseñas de la literatura utilizada, las hipótesis planteadas, los métodos que se emplearon para someterlas a pruebas, los resultados que se obtienen, las conclusiones generadas y la bibliografía.

2.3 La Prosopografía

2.3.1 Definición

Dentro del campo de la investigación histórica uno de los métodos más utilizados sobre todo en la historia política es la prosopografía, en este sentido Stone, la define como:

Investigación retrospectiva de las características comunes a un grupo de protagonistas históricos, mediante un estudio colectivo de sus vidas. El método que se emplea es establecer un universo de análisis, y luego formular una serie uniforme de preguntas acerca del nacimiento y la muerte, el matrimonio y la familia, los orígenes sociales y la posición económica heredada, el lugar de residencia, la educación, el monto y la fuente de la riqueza personal, la ocupación, la religión, la experiencia en cuanto a un oficio, etcétera. Posteriormente, los diversos tipos de información sobre los individuos comprendidos en este universo, se combinan y se yuxtaponen, y se examinan para buscar variables significativas. Se evalúan con respecto a sus correlaciones internas y a sus correlaciones con otras formas de conducta o de acción [1]

Con este método los historiadores, principalmente aquellos especializados en historia política pueden contestar preguntas tales como ¿Por qué determinado grupo político defiende tales o cuales ideas, modos de hacer la política o intereses económicos? ¿Por qué las instituciones políticas asumen ciertas características bajo el influjo de la acción

[1] Lawrence Stone, El pasado y el presente, México, Fondo de Cultura Económica, 1986, p. 61.

de determinados grupos políticos? ¿Cómo la posición social, la riqueza personal y la actividad profesional determinan los comportamientos de los actores políticos? (SORDO R. 1998)

Entre los diferentes grupos de élite que se estudian mediante el auxilio de la prosopografía se encuentran:

Congresos
Legislaturas locales
Agiotistas
Empresarios,
Nobles
Intelectuales
La burocracia

2.3.2 Origen y Evolución

La prosopografía surge en Alemania a principios del siglo XX, elaborada como un método, por eruditos alemanes especialistas en antigüedad clásica y el Estado Romano, difundiéndose posteriormente en Inglaterra, y de ahí en Francia e Italia.(ROUSSEAU, I, 1990)

A partir de entonces se utiliza para elaborar una biografía colectiva partiendo de una colección de personas que comparten ya sea una función, una actividad, un estatuto. Se le da mucha importancia a las relaciones entre los individuos, ya sea a través de cargos, puestos, familia, negocios, política, etc. Se estudian también los orígenes sociales, regionales, educativos y profesionales, permitiendo seguir el ascenso, lucha, reproducción de las diferentes de los grupos de alguna élite.

El estudio y empleo de la prosopografía se dividió entonces en 2 escuelas, la elitista que se dedica a estudiar grupos reducidos pertenecientes a la élite de poder cuyo comportamiento era significante para la sociedad y la escuela de masas que, desde un análisis más estadístico estudiaba a la sociedad en conjunto al grueso de la masa y sus integrantes.

La prosopografía fue evolucionando en su enfoque, en primera instancia, se le daba más importancia a los aspectos económicos de los individuos pertenecientes al grupo, para, de esta forma, tratar de explicar su comportamiento en función de sus intereses materiales, con el análisis de estos trabajos se detectaron errores y limitaciones por lo que al perder confiabilidad se fue dando mayor importancia a otros aspectos como el tratar de detectar las causas que orillaban a ciertas instituciones o individuos a tomar decisiones por lo que se profundizo en buscar datos familiares y sociales. Finalmente y con la madurez de la metodología se emplea para descubrir los vínculos familiares, sociales económicos o políticos que mantienen o explican la cohesión y formación de cierto grupo.

2.3.3 Ventajas

-Permite reconstruir evoluciones temporales de variables propias o de interés para construir tipos.

-Mediante el análisis de datos familiares permite una reconstrucción genealógica de las familias en el poder o influyentes dentro de una sociedad.

-Permite conocer los mecanismos que influyen en la conformación o integración y reproducción de una élite.

-Permite construir correlaciones y vínculos de diferente tipo entre los miembros del universo y reconocer los conjuntos de redes sociales.

-Fomenta la detección de pautas, estadísticas y tendencias en el comportamiento de ciertos grupos de élite.

-Permite establecer la diversificación de comportamiento.

2.3.4 Desventajas

- La documentación no es uniforme, mientras se encuentran grandes volúmenes de información para unos cuantos individuos, para los otros es mínima, en caso de carecer de fuentes primarias seriadas.
- El método es funcional en grupos pequeños, pero se hace impreciso conforme aumenta la cantidad de individuos o el periodo de tiempo.
- Es muy eficaz para resolver problemas específicos, pero se complica cuando el número de variables aumenta.

2.3.5 Proceso

1. El historiador formulará las diferentes interrogantes que desea responderse respecto al origen y comportamiento económico, político, familiar o personal de los individuos del grupo de estudio.
2. Delimitación temporal y espacial del Grupo o élite. Elección de los individuos que conforman el grupo en cuestión y del contexto temporal en el que se ubican sus interrogantes.
3. Identificación de las variables y datos relevantes para la investigación. Planteamiento del conjunto de datos como Datos Personales (Nombre, Estado Civil, lugar de nacimiento, Edad, ocupación) , Datos Familiares(Nombres de Esposa, hijos, padres, hermanos, etc.), Datos Culturales (religión, publicaciones, filiación a clubs sociales, creencias, etc.) Datos Económicos (capital, negocios, ocupaciones, deudas, herederos, acreedores, Propiedades, etc.), Datos políticos (corrientes política, ideologías, manifiestos).
4. Búsqueda e identificación de las diversas fuentes. Ubicación de las fuentes (bibliotecas, hemerotecas, archivos generales, etc.) y tipos de documentos

(libros, actas, oficios, ensayos, periódicos, etc.) que solventarán las variables definidas en el paso anterior.

5. Selección de documentos. Acopio de los documentos o actas, fotocopiado o transcripción.
6. Acopio e integración de los datos. Captura y organización de los datos ya sea en algún medio físico como carpetas, archiveros y demás o en medios electrónicos, escaneo o computadoras.
7. Depuración de los datos. Revisar que los datos no contengan errores de captura, o interpretación, buscar vacíos o elementos redundantes.
8. Codificación de los datos. Numerar y codificar a los sujetos de estudio, así como sus atributos para generar colecciones y facilitar la cuantificación serialización e identificación
9. Análisis de los datos en conjunto. Estudio de los datos a partir del conocimiento previo del historiador, búsqueda de pautas, tendencias y análisis de las relaciones interpersonales de los individuos.
10. Interpretación de la información obtenida por el análisis. Contestar y evaluar si los datos obtenidas en el análisis y procesamiento de los datos contesta o no las interrogantes planeadas y cubre las expectativas de nuestro caso de estudio.
11. Redacción de informes y estadísticas. Integración de un informe auxiliado con datos duros, tablas estadísticas, gráficos y demás.

3.1 Introducción

La actual evolución informática presente en los tiempos modernos, ha determinado el camino lógico que seguirán muchas disciplinas del conocimiento humano en auxilio de su desarrollo, Tal es el caso de las ciencias sociales y humanísticas, particularmente la investigación histórica.

Los grandes y diversos volúmenes de información generados por los investigadores de la historia, y la facilidad que actualmente tienen de almacenarlos en computadoras o medios digitales para su posterior análisis, clasificación o reutilización hacen preciso la sistematización, formalización e integración de metodologías, herramientas informáticas y algoritmos.

Los sistemas de información son definidos como el conjunto de soluciones integradas y concretas a problemas específicos de una actividad humana que precise el procesamiento de grandes volúmenes de información, están enfocados a realizar 4 funciones básicas: entrada, almacenamiento, procesamiento y salida de información. (GARCÍA, F.J. 2003)

Entrada de Información: Es el proceso mediante el cual el Sistema de Información toma los datos que requiere para procesar la información. Las entradas pueden ser manuales o automáticas. Las manuales son aquellas que se proporcionan en forma directa por el usuario, mientras que las automáticas son datos o información que provienen o son tomados de otros sistemas o módulos. Esto último se denomina interfaces automáticas.

Las unidades típicas de entrada de datos a las computadoras son las terminales, las cintas magnéticas, las unidades de disco compacto, los códigos de barras, los escáneres, la voz, los monitores sensibles al tacto, el teclado y el mouse, entre otras.

Almacenamiento de información: El almacenamiento es una de las actividades o capacidades más importantes que tiene una computadora, ya que a través de esta propiedad el sistema puede recordar la información guardada en la sección o proceso anterior. Esta información suele ser almacenada en estructuras de información denominadas archivos. La unidad típica de almacenamiento son los discos magnéticos o discos duros, los discos flexibles o diskettes y los discos compactos (CD-ROM).

Procesamiento de Información: Es la capacidad del Sistema de Información para efectuar cálculos de acuerdo con una secuencia de operaciones preestablecida. Estos cálculos pueden efectuarse con datos introducidos recientemente en el sistema o bien con datos que están almacenados. Esta característica de los sistemas permite la transformación de datos fuente en información que puede ser utilizada para la toma de decisiones, lo que hace posible, entre otras cosas, que un tomador de decisiones genere una proyección financiera a partir de los datos que contiene un estado de resultados o un balance general de un año base.

Salida de Información: La salida es la capacidad de un Sistema de Información para sacar la información procesada o bien datos de entrada al exterior. Las unidades típicas de salida son las impresoras, terminales, diskettes, cintas magnéticas, la voz, los graficadores y los plotters, entre otros. Es importante aclarar que la salida de un Sistema de Información puede constituir la entrada a otro Sistema de Información o módulo. En este caso, también existe una interface automática de salida.

3.2 Clasificación de los Sistemas de Información

Según la función a la que vayan destinados o el tipo de usuario final del mismo, los Sistemas de Información pueden clasificarse en:

* Sistema de procesamiento de transacciones (TPS).- Gestiona la información referente a las transacciones producidas en una empresa u organización.
* Sistemas de información gerencial (MIS).- Orientados a solucionar problemas empresariales en general.
* Sistemas de soporte a decisiones (DSS).- Herramienta para realizar el análisis de las diferentes variables de negocio con la finalidad de apoyar el proceso de toma de decisiones.
* Sistemas de información ejecutiva (EIS).- Herramienta orientada a usuarios de nivel gerencial, que permite monitorizar el estado de las variables de un área o unidad de la empresa a partir de información interna y externa a la misma.
* Sistemas de automatización de oficinas (OAS).- Aplicaciones destinadas a ayudar al trabajo diario del administrativo de una empresa u organización.
* Sistema experto (SE).- Emulan el comportamiento de un experto en un dominio concreto.

Debido a que esta clasificación obedece principalmente a necesidades empresariales o del sector negocios, no se encuentra dentro de ellos perfectamente definido un tipo de Sistemas de Información que englobe a todos aquellos orientados a las ciencias humanísticas y sociales, sistemas que van cobrando mucha relevancia en los últimos años y que han sido producto de un desarrollo acelerado por parte de investigadores y maestros involucrados en estas áreas.

Tal es el caso, entre otros, el de los Sistemas de Información Geográfica que en los últimos años ha explotado y se ha desarrollado a pasos agigantados dentro del sector público y social y que se ha convertido en una herramienta necesaria para diversas dependencias gubernamentales, universidades e institutos.

3.3 Definición

Dentro de este tipo ha surgido el concepto de Sistemas de Información Histórica debido a la necesidad de procesamiento de grandes volúmenes de información almacenada referente al estudio del pasado. Tal concepto no ha despegado totalmente , comparado con el de Sistemas de Información Geográfica sin embargo la aplicación de los sistemas informáticos como auxiliares en la investigación histórica ha avanzado enormemente.

Una de las posibles causas por las que los SIH no han despegado la plantea Francisco Javier García Marco Investigador-Profesor de la Universidad de Zaragoza, España:

> *Entre las muchas causas posibles de esta situación, interesa resaltar una explicación que proviene de la propia intensión del concepto SIH. Un SIH es, efectivamente, un sistema; en definitiva, una realidad integrada y un proyecto colaborativo. Por el contrario, la tradición de la investigación en Humanidades es, desde sus mismos orígenes, todo lo individualista que puede serlo sin dejar de ser compatible con el concepto de comunidad científica, debido en parte a la propia naturaleza de la relación entre sujeto y objeto en estas ciencias. En Humanidades, salvando excepciones, la investigación se basa sobre todo en esfuerzos personales; eso sí, organizados en corrientes filosófico-científicas. Raramente, esa interacción débil que caracteriza a los humanistas y a muchos científicos sociales se concreta en proyectos colectivos organizados. Sin embargo, los investigadores en Humanidades han estado históricamente de acuerdo en colaborar en determinadas tareas.*[1]

El SIH se orienta a facilitar el desarrollo de una comunidad científica en el ámbito de las ciencias sociales, por medio de la objetivación, estructuración, normalización y depuración compartida de los datos y conocimientos históricos. (fig. 1)

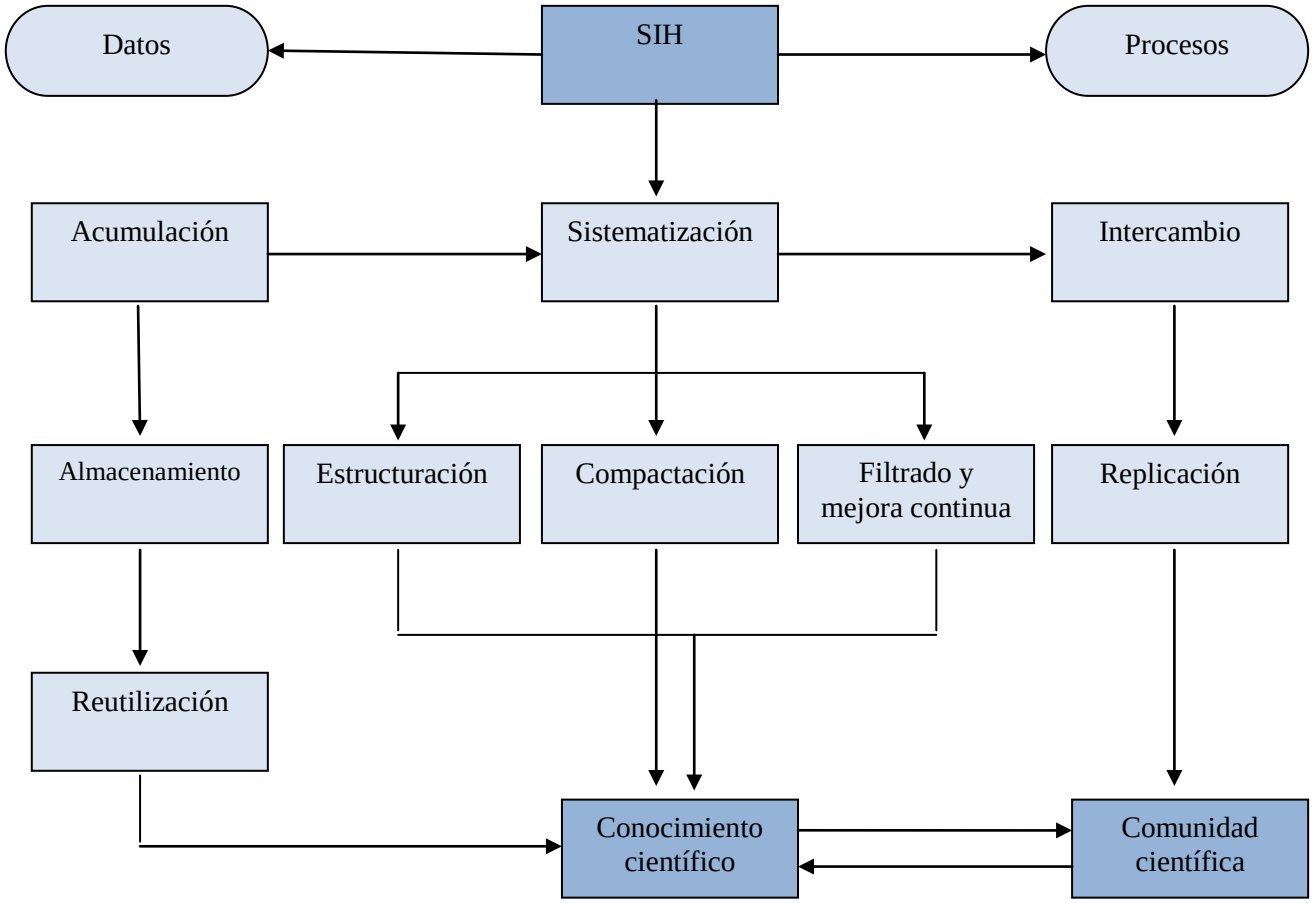


Figura 1: Integración de los SIH en la actividad científica

[1] García Marco, Francisco Javier. Sistemas de información histórica: la documentación al servicio del pasado. En: Bibliodoc 2001/2002: Anuario de Biblioteconomía, Documentación e Información / María Eulalia Fuentes i Pujol (Dir.). Barcelona: Colegio Oficial de bibliotecaris-Documentalistes de Catalunya, 2003. p. 76

3.4 Ventajas de un SIH.

- Investigación más eficaz y rentable
- Diseminación de resultados más eficaz
- Instituciones de investigación más fuertes
- Mejor comprobación de las hipótesis historiográficas entre colegas
- Disponibilidad universal de los documentos, datos y procedimientos
- Mejor integración entre archiveros, informáticos, documentalistas, bibliotecarios, e historiadores

Podemos entender que un Sistema de Información Histórica va más allá de una simple colección de datos y fuentes ordenados y clasificados, lo anterior es parte de él, sin embargo también es la integración de un sistema informático de fuentes primarias y secundarias, datos y procedimientos que obedecen a teorías y metodologías compartidas por una comunidad de investigadores. No es simplemente aplicar herramientas informáticas a la investigación histórica, es la aplicación sistemática que parte de instituciones y grupos de investigación que comparten sus datos y procedimientos aunque no estén de acuerdo entre ellos de su uso. Se trata de integrar Fuentes textuales primarias –documentos de archivo- y secundarias -historiografía- con Bases de Datos.

3.5 Tipos de SIH

Se pueden clasificar hasta cinco tipos de sistemas de información histórica:

1. Sistemas de información histórica orientados al control, análisis y difusión de datos de valor primordialmente legal, patrimonial, económico, político, etc. En definitiva, a esa franja de información que es a la vez permanente y operativa.
2. Sistemas de información histórica orientados al control, análisis y difusión del patrimonio documental, que quizás habría que redenominar patrimonio informacional que cada vez se crea y utiliza más información en bases de datos.
3. Sistemas de información histórica orientados al control, análisis y difusión del patrimonio histórico-artístico y del natural, que comparten el interés gerencial –la orientación al control– con el servicio a la investigación y a las actividades turísticas. De hecho, términos como “heritage information system” o “cultural information system” son cada vez más frecuentes en el ámbito anglosajón, y contrastan con el uso marginal que se hace de la expresión SIH.
4. Sistemas de información histórica orientados a la investigación sobre el pasado, como los sistemas de información arqueológica, etc.

5. En realidad, toda acumulación de datos termina generando sistemas de información histórica ligados a los sistemas de información para la gestión empresarial, asociativa o pública que dieron origen a los mismos, y que, eventualmente, pueden terminar formando parte de un SIH orientado a la conservación y difusión del patrimonio informacional (Fig. 2). A este tipo de SIH se les podría denominar derivados, pues resultan de la acumulación normal de información en un sistema de información operacional.

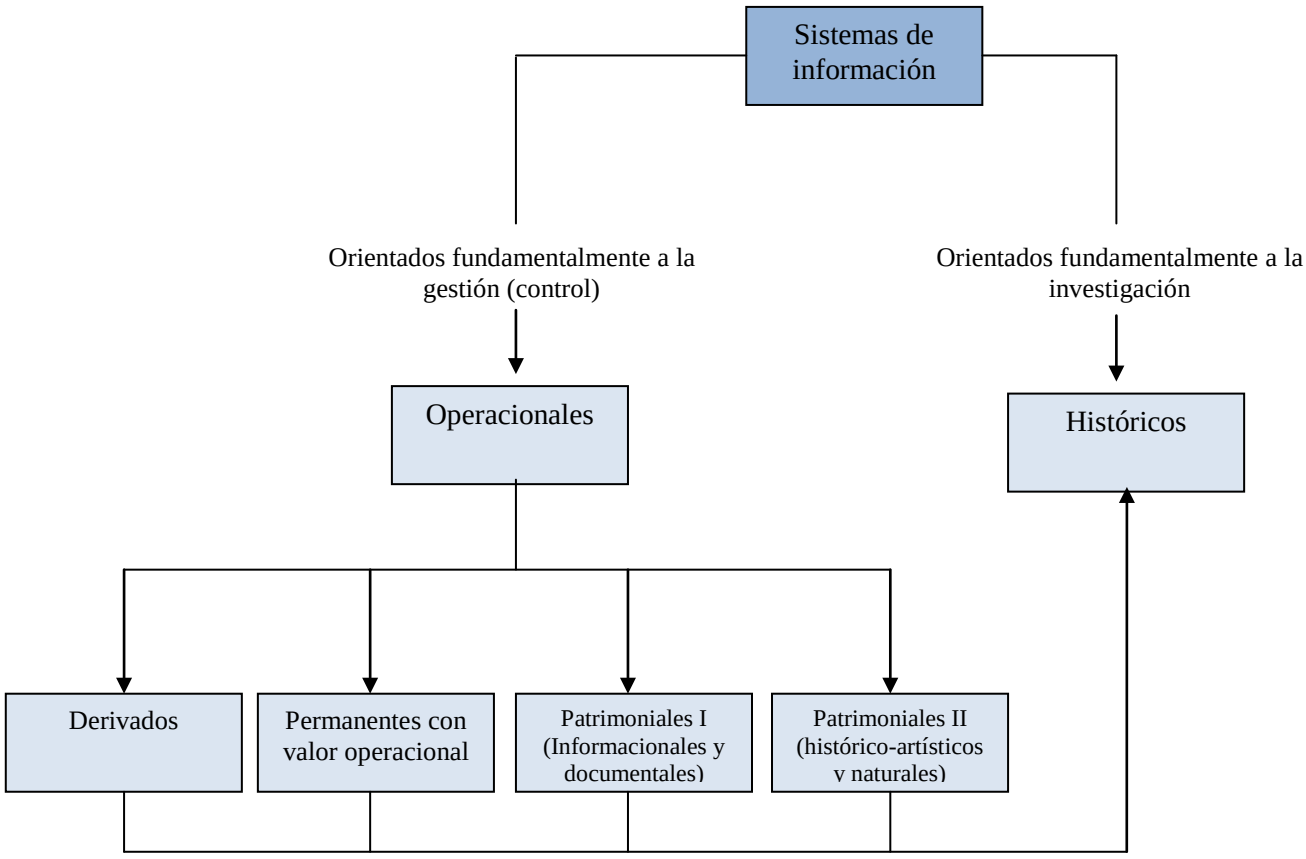


Figura 2: Clasificación de los Sistemas de Información a partir de los SIH

De lo anterior se puede entender en una definición más formal a los Sistemas de Información Histórica como organizaciones humanas que diseñan, administran y utilizan sistemas automatizados para almacenar, tratar, y recuperar información histórica representada en un conjunto de bases de datos integradas que evolucionan continuamente y son utilizados con fines de investigación.

3.6 Elementos principales de un SIH (fig.3)

- Agentes humanos organizados.
- Objetivos de investigación.

- Sistemas informáticos.
- Bases de datos (que son la plasmación concreta de la información histórica
- Procedimientos formales de tratamiento de la información (metodologías).

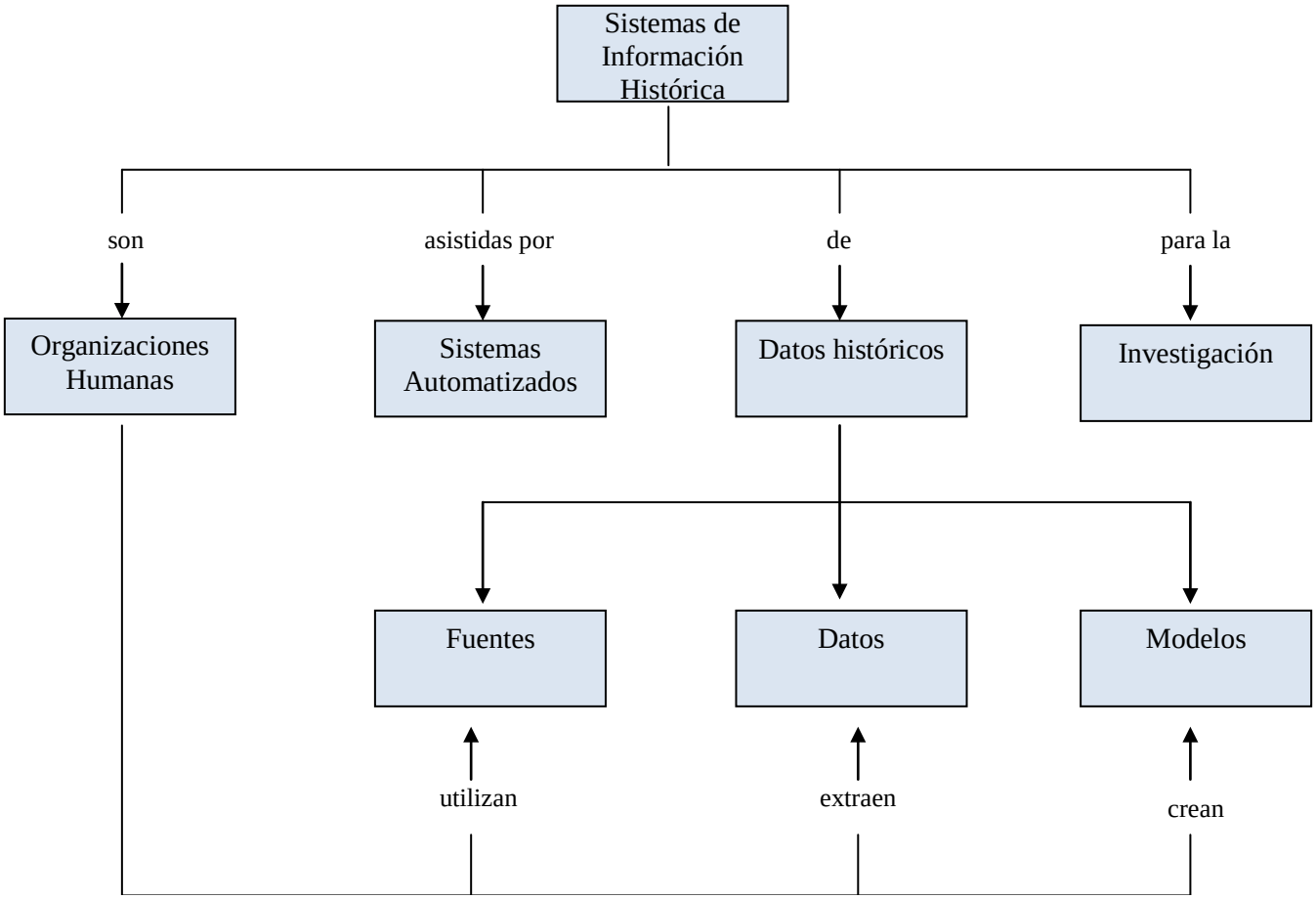


Figura 3: Elementos de un SIH

Agentes humanos organizados

El subsistema humano es el principal de todos, pues es el depositario de los valores y objetivos del sistema, o –expresándolo en términos de la propia definición de sistema– de los fines en torno a los cuales se organiza dinámicamente el resto de los elementos que lo forman.

El subsistema humano se estructura siempre en algún tipo de organización, ya sea jerárquica o asociativa. La experiencia de los proyectos exitosos en el mundo confirma que los mejores SIH están soportados por una red de alianzas entre instituciones de investigación y de custodia vertebrados por alguna organización especialmente dedicada, que se ha especializado debido a las demandas de integración de datos y procedimientos que sus actividades plantean.

Un análisis más estricto permitiría deslindar dentro de lo que hemos denominado el subsistema humano otros dos: el organizacional y el del personal.

En cualquier caso, el desarrollo de SIH tiene como tarea prioritaria y aun previa el desarrollo de estas redes, alianzas, instituciones, etc., en las cuales, por cierto, las unidades de información y documentación y sus profesionales desarrollan siempre un papel decisivo.

Objetivos de investigación

Los SIH no implican tan sólo la aplicación de herramientas informáticas al control y explotación de la información sobre el pasado, sino que suponen un nuevo concepto en la organización del trabajo histórico, que implica la integración y normalización de datos, fuentes, procedimientos, metodologías y teorías.

Su objetivo es integrar, en un mismo sistema, datos, tratamientos, representaciones documentales, documentos y conocimientos, perspectivas personales y disciplinares, de tal manera que resulte posible un trabajo histórico más acumulativo, compartido, discutido y, en definitiva, más controlado por el método científico.

Normas

El sistema de valores y objetivos suele en la práctica terminar independizándose del subsistema humano para, junto con otros elementos, configurar una especie de superestructura, que termina tomando un cierto control sobre los componentes humanos.

Se trata del sistema filosófico-normativo, esto es, de las declaraciones de principios y de las normas de estructuración, interacción, producción y reproducción del sistema.

En el tema concreto que nos ocupa, cabe destacar la importancia de las políticas de cooperación y de las normas de intercambio de datos, y, especialmente dentro de estas últimas de las referentes a las estructuras de datos y metadatos.

El subsistema informático

Por lo que se refiere a los sistemas de automatización de información –esto es, sistemas informáticos– señalar tan sólo que constan de dos partes bien diferenciadas.

Por una parte, su infraestructura que suele sustentarse en dispositivos y programas estándar para la entrada, el procesamiento, el almacenamiento y la difusión de los datos.
Por otra parte, sobre esa infraestructura se programa todo un conjunto de procedimientos específicos de tratamiento de la información que pueden ir de lo más

simple –gestión del almacenamiento y recuperación de los datos–, a análisis complejos –por ejemplo, de tipo estadístico o de representación de reglas.

El subsistema de los datos: la información

El quinto elemento de un SIH –éste sí absolutamente característico– son precisamente los datos que éste almacena y procesa.

El aspecto definitorio de dichos datos es su carácter histórico. Se trata de datos que no se están utilizando en las actividades ordinarias del funcionamiento personal, organizacional o social; es decir, que no tienen un valor operativo –o, al menos, que no lo tienen de forma continua, urgente y frecuente.

Otra propiedad muy importante de los datos representados en los SIH que debe ser destacada es su carácter complejo y estratigráfico. De hecho, dichos datos se integran como mínimo en cinco niveles de información diferentes:

Los metadatos, que ayudan a describir coherentemente todos los otros niveles de entidades de acuerdo con la ontología –o meta-ontología, si ha sido necesario compatibilizar diferentes ontologías– que comparte la comunidad científica que utiliza el SIH.

Documentos, que se concretan en facsímiles digitales de originales físicos y documentos digitales originales, que se preservarán lógicamente en un archivo físico o digital permanente, con el que el SIH mantendrá estrechas relaciones institucionales, humanas y, claro está, de compatibilidad informática.

Representaciones documentales
• Representaciones bibliográficas (referencias de literatura científica, facsímiles de fuentes documentales primarias, reproducciones gráficas de objetos y entornos, etc.).
• Representaciones archivísticas (fuentes documentales primarias no bibliográficas).
• Representaciones museográficas y patrimoniales (objetos y entornos).

Datos analíticos históricos.
Relaciones entre datos, bien sean expresadas a través de estructuras de almacenamiento o bien mediante un programa (por ejemplo, en SQL).

3.7 Conclusiones

La última aportación del mundo de la informática y más concretamente de las avanzadas tecnologías de la información, al campo de la investigación histórica la constituyen los denominados sistemas de Información Histórica, Los SIH nacen con la pretensión de integrar en un solo instrumento todas las posibilidades de gestión documental de las bases de datos perfiladas para el trabajo de investigación histórica.

Un SIH puede definirse como un sistema automatizado que integra un conjunto de bases de datos y de procedimientos formales diseñados y mantenidos para almacenar tratar y recuperar información histórica.

Un SIH debe ser capaz de almacenar fuentes (tanto sus referencias como sus representaciones y reproducciones) referencias bibliográficas y trabajos de investigación, ya sea que se presenten en forma de datos, textos, de gráficos o de procedimientos.

El aspecto clave de un SIH es precisamente la interface entre todos estos aspectos y su integración dinámica. Esta integración es la que marca la diferencia entre un mero archivo electrónico (un conjunto de documentos, gráficos y bases de datos sobre un determinado tema, pero sin conexión dinámica entre ellos) y un auténtico sistema de información

Estos sistemas de información necesitan de una puesta al día continua y es absolutamente necesario su integración en modernas redes científicas de investigación que permitan su total implantación y desarrollo. En este sentido los SIH no presentan en la actualidad el mismo grado de desarrollo que otros instrumentos similares que parecen vivir un momento de plena expansión de áreas de estudio como la geografía, la administración, los archivos, etc. Una de las causas de que los SIH no estén tan implantados es por el poco interés de los profesionales de la Historia a la hora de manejar y desenvolverse en el terreno de las tecnologías de la información.

Los principales objetivos entonces de un SIH son:

Estructurar información en forma de bases de datos cuantitativo, textual factual y multimedia.

Evitan la duplicación de tareas de recopilación y organización documental en cuanto la información estructurada es accesible, conectable e incorporable a otros sistemas de información histórica.

Propician la especialización de la documentación histórica por parte de las instituciones responsables de la construcción de tales sistemas.

Contribuyen a normalizar la ciencia histórica y su crecimiento en cuanto a la relación entre la fuente histórica, la historiografía y el trabajo del investigador actual.

Tratan de integrar todos los elementos referidos a las fuentes y a los resultados de las investigaciones, lo que incluye la normalización de las operaciones y de la terminología.

4. Minería de datos

4.1 Definición

El término Minería de datos como parte del proceso de descubrimiento de conocimiento en bases de datos (KDD), que se tratará más adelante, está claramente definido de manera breve y concisa como "Extracción de información útil a partir de grandes cantidades de datos", sin embargo cabe hacer mención y se procura enriquecer y detallar tal concepto dada la inquietud del presente trabajo por vincularlo con el proceso de investigación histórica y su futura aplicación dentro de las ciencias sociales. (HERNÁNDEZ, J. 2004)

La extracción no trivial de información implícita, que previamente es desconocida y que potencialmente es útil para la obtención de conocimiento nos permitirá tomar decisiones y, en el caso de la investigación histórica, formular posibles hipótesis de hechos sobre el pasado.

La tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos haciendo uso de diferentes tecnologías que resuelven problemas típicos de agrupamiento automático, clasificación, asociación de atributos y detección de patrones secuenciales, proceso que para ser efectivo, deberá ser automático o semi-automático-asistido- y el uso de estos modelos deberá generar siempre una utilidad o beneficio, en este caso a la investigación histórica.

Es importante establecer la diferencia entre un análisis estadístico de datos y un proceso de minería de datos, que aunque aparentemente ambos persiguen el mismo fin -obtener información útil analizando los datos- en el caso de los análisis estadísticos la obtención del conocimiento y las posibles afirmaciones generadas después de tal proceso conlleva a conocer a priori las relaciones y los argumentos que harán válida tal o cual afirmación, en el caso de la minería de datos no es así puesto que obtendremos conocimiento y afirmaciones que los datos arrojarán por si mismos, es decir en el primer caso obtendremos respuestas o afirmaciones claras a preguntas específicas previamente definidas, mientras en el segundo se obtendrán afirmaciones arrojadas por los datos que no necesariamente responderán a preguntas estructuradas.

4.2 **Relación con otras disciplinas**

Dado que la minería de datos surge como un campo multidisciplinar en relación con otras técnicas (fig. 4.1) con un crecimiento ya sea paralelo o como prolongación de las mismas, la investigación y los avances que se tienen en este campo está directamente ligado al avance de las áreas relacionadas.

Dentro de las áreas más destacadas que influyen en la minería de datos tenemos:

Las bases de datos: los almacenes de datos (data warehouses) y el procesamiento analítico en línea (OLAP) tienen una relación muy cercana con las técnicas de minería de datos.

La recuperación de información (information Retrieval) que es el proceso de encontrar información a partir de datos textuales que históricamente se ha basado en el uso de bibliotecas que en la actualidad son digitales. Como ejemplo típico tenemos la tarea de recuperar documentos de una colección a partir de palabras clave. Esto genera procesos de clasificación de documentos en función de las palabras clave utilizando medidas de similitud entre los documentos y las palabras clave -manejo de cercanía de vectores- . Medidas que se utilizan en algunas aplicaciones generales de minería de datos.

La estadística: La mayoría de los conceptos, algoritmos y técnicas de minería de datos han sido proporcionados por la estadística, como la varianza, la media, las distribuciones, el análisis univariable, el análisis multivariable, la regresión lineal, la no lineal, la teoría del muestreo, la validación cruzada, las técnicas bayesianas, la modelización paramétrica y la no paramétrica, entre otras muchas más, esta es una de las razones por las que en algunos casos se confunde la aplicación de estadística con la minería de datos.

El aprendizaje automático: un área de la inteligencia artificial que desarrolla algoritmos o programas capaces de aprender que constituye el centro mismo del análisis inteligente de los datos, el aprendizaje automático y la minería de datos siguen la misma técnica, la máquina aprende un modelo a partir de ejemplos y posteriormente utiliza ese modelo para resolver problemas u obtener resultados.

La visualización de los datos: Fundamental es esta área pues permite al usuario descubrir, intuir y entender patrones que le serían muy difíciles de apreciar puestos en términos matemáticos o textuales arrojados por la máquina, por ello es preciso contar con gráficas(diagramas de barras, gráficas de dispersión, histogramas, etc), icónicas, las basadas en pixeles, las jerárquicas, entre otras.

Computación Paralela y distribuida: debido a los grandes volúmenes de información y necesidades de procesamiento actualmente las bases de datos se procesan en paralelo, o distribuido dado que las mismas bodegas de datos se encuentran distribuidas por lo que la minería de datos debe también aplicarse bajo este tenor, si el caso así lo amerita.

Otras disciplinas: La relación de la minería de datos con otras disciplinas depende mucho del tipo de los datos analizar, como puede ser el manejo del lenguaje natural, el análisis de imágenes, el procesamiento de señales, los gráficos por computadora, etc.

Uno de los objetivos del presente trabajo es precisamente relacionar a la historiografía, las técnicas de investigación histórica (en este caso la prosopografía) con la minería de datos.

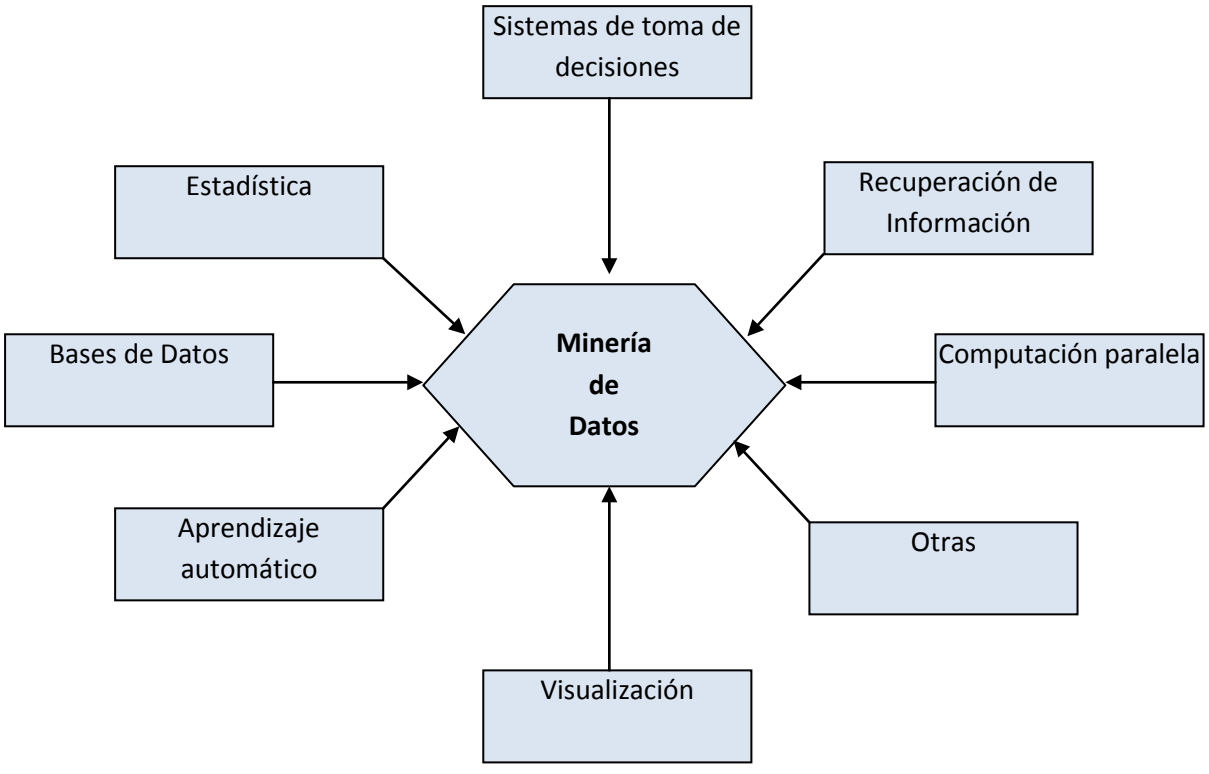


Figura 4.1: Relación de otras disciplinas con la minería de datos

4.3 Descubrimiento de conocimiento en base de datos (KDD)

El Descubrimiento del Conocimiento en Bases de Datos (Knowledge Discovery in DataBases, KDD) está definido como un proceso no trivial de identificar patrones válidos novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos.

El KDD tiene como metas procesar automáticamente grandes cantidades de datos crudos, identificar los patrones más significativos y relevantes, y presentarlos como conocimiento apropiado para satisfacer las metas del usuario.

En muchas bibliografías y trabajos similares un error recurrente es la utilización indistinta de los términos de KDD y Minería de Datos como un mismo concepto, sin tener en cuenta que el KDD es un proceso que consta de varias fases siendo la Minería de Datos una de estas fases. Este proceso genera un conocimiento extraído que debe tener las siguientes propiedades.

- Valido. Los patrones deben seguir siendo precisos aún para datos nuevos que han sido introducidos en la Base de Datos y no sólo para los datos que han sido utilizados para la obtención de dichos patrones.
- Novedoso. Debe aportar información novedosa tanto para el sistema y sobretodo para el usuario.
- Útil. La información obtenida debe conducir a acciones que aporten algún tipo de beneficio al usuario.
- Comprensible. La extracción de patrones que no es comprensible y clara imposibilita su interpretación, revisión, validación y uso en la toma de decisiones. Si la información no es comprensible deja de ser útil por lo tanto no proporciona conocimiento alguno y el proceso fracasa.

Lo anterior muestra que el proceso de KDD (fig. 4.2) es más complejo y además de obtener modelos y patrones (objetivos de la Minería de datos) también plantea el análisis, interpretación y evaluación de los mismos.

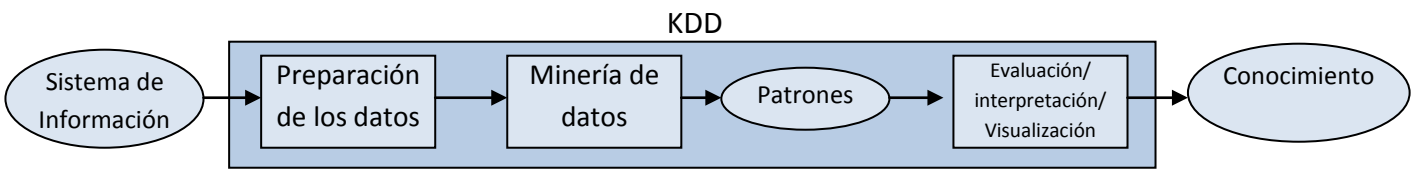


Figura 4.2: Proceso KDD

4.3.1 Componentes del KDD

- **Conocimiento del dominio y preferencias del usuario:** Incluye el diccionario de datos, información adicional de las estructuras de los datos, restricciones entre campos, metas o preferencias del usuario, campos relevantes, listas de clases, jerarquías de generalización, modelos causales o funcionales, etc.

  El objetivo del conocimiento del dominio es orientar y ayudar en la búsqueda de patrones interesantes (aunque a veces puede causar resultados contraproducentes).

  Se tiene que hacer un balance entre eficiencia y completes del conocimiento.

- **Control del descubrimiento:** Toma el conocimiento del dominio, lo interpreta y decide qué hacer (en la mayoría de los sistemas el control lo hace el usuario).
- **Interfaces:** Con la base de datos y con el usuario
- **Foco de atención:** Especifica qué tablas, campos y registros acceder. Tiene que tener mecanismos de selección aleatoria de registros tomando muestras estadísticamente significativas, puede usar predicados para seleccionar un subconjunto de los registros que comparten cierta característica, etc.

  Algunas técnicas para enfocar la atención incluyen:

  - Agregación: junta valores (por ejemplo, los más bajos y los más altos)
  - Partición de datos: en base a valores de atributos (por ejemplo, sólo aquellos datos que tengan ciertos valores)
  - Proyección: ignorar algún(os) atributo(s)

  Partición y proyección implican menos dimensiones. Agregación y proyección implican menos dispersión.

- **Extracción de patrones:** Donde patrón se refiere a cualquier relación entre los elementos de la base de datos. Pueden incluir medidas de incertidumbre. Aquí se aplican una gran cantidad de algoritmos de aprendizaje y estadísticos.
- **Evaluación:** Un patrón es interesante en la medida que sea confiable, novedoso y útil respecto al conocimiento y los objetivos del usuario. La evaluación normalmente se le deja a los algoritmos de extracción de patrones que generalmente están basados en significancia estadística (sin embargo, no es ni debe ser el único criterio).

4.4 Fases del proceso de descubrimiento del conocimiento en bases de datos

El descubrimiento del conocimiento en bases de datos es un proceso que requiere seguir cuidadosamente una serie de pasos que nos ayuden a extraer de nuestras bases de datos en bruto conocimiento y patrones de interés particular.

En primer lugar, antes de usar nuestras bases de datos y comenzar con un largo proceso sobre ellas es necesario contar con un entendimiento claro del dominio de la aplicación, el conocimiento relevante a usar y, por sobre todo, las metas del usuario, esta quizá puede ser una tarea que puede consumir mucho tiempo. Una vez que se han planteado perfectamente las metas y los posibles resultados que debemos obtener es necesario seleccionar un conjunto o un subconjunto de bases de datos, seleccionar y enfocar la búsqueda en subconjuntos de variables, y seleccionar muestras de datos en donde realizar el proceso de descubrimiento.(HAN, J. 2006)

La limpieza y pre-procesamiento de nuestros datos es el siguiente paso, para lo cual se debe diseñar una estrategia adecuada para manejar ruido, valores incompletos, secuencias de tiempo, casos extremos, etc. Después el proceso continúa seleccionando la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, regresión, etc.

Se seleccionan los algoritmos a utilizar, se transforman los datos al formato requerido por el algoritmos específico de minería de datos, es el momento de llevar a cabo el proceso de minería de datos buscando patrones que pueden expresarse como un modelo o simplemente que expresen dependencias de los datos; el modelo encontrado depende de su función y de su forma de representarlo, se tiene que especificar un criterio de preferencia para seleccionar un modelo dentro de un conjunto posible de modelos y se debe especificarla estrategia de búsqueda a utilizar.

El siguiente paso consiste en interpretar los resultados y posiblemente regresar a los pasos anteriores lo cual puede significar repetir el proceso con otros datos, otros algoritmos, otras metas y otras estrategias, este es un paso crucial en donde se requiere tener conocimiento del dominio. La interpretación puede beneficiarse de procesos de visualización, y sirve también para borrar patrones redundantes o irrelevantes.

El conocimiento se obtiene para realizar acciones, ya sea incorporándolo dentro de un sistema de desempeño o simplemente para almacenarlo y reportarlo a las personas interesadas. En este sentido, el proceso de extracción del conocimiento implica un proceso iterativo que involucra interacciones complejas entre herramientas heterogéneas.

Los pasos mencionados anteriormente emplean distintas técnicas de distintas disciplinas. En la figura 4.3 se plantea esta relación.

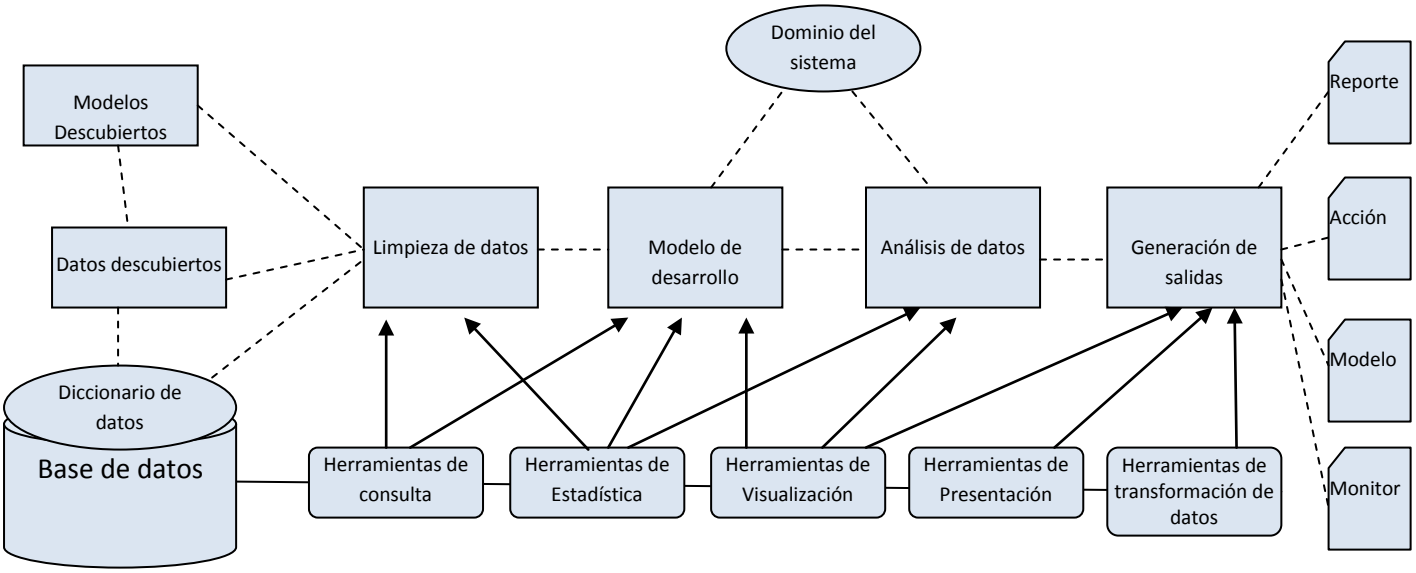


Figura 4.3: Disciplinas involucradas en el proceso KDD

4.4.1 Fase de selección, limpieza y transformación

La calidad del conocimiento obtenido después de la minería de datos no sólo depende de la correcta selección del algoritmo a utilizar o de la claridad en el planteamiento del resultado deseado, también depende, en gran medida, de la calidad de los datos minados. El siguiente paso en el proceso de KDD es la selección y adecuada preparación de los datos que serán minados. Lo anterior se hace preciso dada la posibilidad de que existan datos dentro de los conjuntos seleccionados que sean irrelevantes o innecesarios.

Uno de los problemas más comunes que afectan la calidad de los datos es la presencia de valores que no se ajustan al comportamiento general de los datos llamados outliers. Estos valores pueden representar errores en los datos o simplemente ser valores que son correctos que son diferentes a los demás. Existen algoritmos de Minería de datos que ignoran estos datos, o los consideran excepciones o ruido, sin embargo otros son muy sensibles y pueden alterar sustancialmente los resultados. No siempre es conveniente eliminarlos, ya que en algunos casos estas excepciones son las que revelan tendencias que pueden ser de gran utilidad. Algunas acciones que se pueden tomar para prevenir los outliers son:

Ignorar: algunos algoritmos son robustos a datos anómalos (p.ej. árboles)

Filtrar (eliminar o reemplazar) la columna: solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad. Preferible a eliminar la columna es reemplazarla por una columna discreta diciendo si el valor era normal u outlier (por encima o por debajo).

Filtrar la fila: puede sesgar los datos, porque muchas veces las causas de un dato erróneo están relacionadas con casos o tipos especiales.

• reemplazar el valor: por el valor 'nulo' si el algoritmo lo trata bien o por máximos o mínimos, dependiendo por donde es el outlier, o por medias. A veces se puede predecir a partir de otros datos, utilizando cualquier técnica de ML.

• discretizar: transformar un valor continuo en uno discreto (p.ej. muy alto, alto, medio, bajo, muy bajo) hace que los outliers caigan en 'muy alto' o 'muy bajo' sin mayores problemas.

El que existan datos faltantes o perdidos dentro de una base de datos, es también un problema que puede conducir a resultados poco precisos o equívocos. Sin embargo en el tratamiento de este problema es necesario hacer un análisis de la causa por la que estos datos falten. A veces es importante examinar las razones tras datos faltantes y actuar en consecuencia:

Algunos valores faltantes expresan características relevantes

Valores no existentes: muchos valores faltantes existen en la realidad, pero otros no

Datos incompletos: si los datos vienen de fuentes diferentes, al combinarlos se suele hacer la unión y no la intersección de campos, con lo que muchos datos faltantes representan que esas tuplas vienen de una/s fuente/s diferente/s al resto.

Posibles acciones ante datos faltantes:

Filtrar la fila: claramente sesga los datos, porque muchas veces las causas de un dato faltante están relacionadas con casos o tipos especiales.

Reemplazar el valor: por medias. A veces se puede predecir a partir de otros datos, utilizando cualquier técnica de ML.

Segmentar: se segmentan las tuplas por los valores que tienen disponibles. Se obtienen modelos diferentes para cada segmento y luego se combinan.

Modificar la política de calidad de datos y esperar hasta que los datos faltantes estén disponibles.

Aunado a lo anterior, es necesario también poder proporcionar a los métodos de minería de datos el subconjunto de datos más adecuado par lo que se requiere seleccionar los datos apropiados. La selección de atributos relevantes es un preprocesamiento muy importante ya que es crucial que los atributos utilizados sean relevantes para la tarea de minería de datos.

Otra tarea de preparación de los daros es la construcción de atributos, la cual consiste en construir automáticamente nuevos atributos aplicando alguna operación o función a los atributos originales con objeto de que estos atributos hagan más fácil el proceso de minería.

El tipo de los datos también puede modificarse para adecuarlos al tipo requerido por los algoritmos de minería de datos. A continuación se analizan un par de técnicas frecuentemente utilizadas para la transformación de campos

Numerización / Etiquetado

• Ventajas: Se reduce espacio. Se pueden utilizar técnicas más simples.

• Desventajas: Se necesita meta-información para distinguir los datos inicialmente no numéricos (la cantidad no es relevante) de los inicialmente numéricos (la cantidad es relevante: precios, unidades, etc.)

Discretización:

• Ventajas: Se reduce espacio. Se pueden utilizar árboles de decisión y construir reglas discretas.

• Desventajas: Una mala discretización puede invalidar los resultados.

4.4.2 Fase de Minería de datos

Esta fase, es considerada la más trascendente del KDD, su objetivo, ya mencionado, es producir un conocimiento que pueda ser utilizado por un usuario final. Esto se realiza construyendo un modelo basado en los datos recopilados para este propósito. El modelo es una descripción de los patrones y relaciones existentes entre los datos que permitan al usuario realizar predicciones, para entender los datos o para explicar situaciones pasadas. Para tal fin es preciso tomar en cuenta una serie de decisiones antes de comenzar con el proceso.

-Determinar claramente el tipo de tarea de minería más adecuado

-Elegir el tipo de modelo

-Elegir el algoritmo de minería que resuelva la tarea y obtenga el tipo de modelo que elegimos.

En la minería de datos se distinguen diferentes tipos de tareas, que pueden considerarse como tipos de problemas que pueden resolverse por un determinado algoritmo de minería, lo cual significa que cada tarea cuenta con sus propios requisitos y que el tipo de información obtenida por una puede ser totalmente diferente de la obtenida por otra. Estas tareas pueden ser predictivas o descriptivas. Las tareas predictivas pretenden estimar valores futuros o desconocidos de variables de interés que se denominan variables objetivo. Usando otras variables o campos de la base de datos a las que nos referimos como variables predictivas o independientes. Las tareas descriptivas identifican patrones que explican, resumen o justifican los datos, sirven para analizar las propiedades de los datos examinados, y no para predecir nuevos datos.

Tareas Predictivas

-Clasificación
-Regresión

Tareas Descriptivas

-Agrupamiento
-Reglas de asociación
-Reglas de asociación secuenciales
-Correlaciones

Clasificación: En esta tarea cada instancia o registro de la base de datos pertenece a una clase, que se indica mediante un atributo llamado clase de la instancia, Este atributo puede tomar valores discretos que corresponden a una clase. Los atributos relevantes restantes de aquella instancia se utilizan para predecir la clase. El objetivo es predecir la clase de las nuevas instancias en las que se desconoce esta.

Regresión: Consiste en aprender una función real que asigna a cada instancia u valor real, pero que a diferencia de la clasificación esta predice un valor numérico. El objetivo es minimizar el error que exista entre el valor predicho y el real.

Agrupamiento: Tarea descriptiva que consiste en obtener grupos naturales a partir de de los datos, a diferencia de la clasificación que analiza datos etiquetados con una clase los analiza generando etiquetas. Los datos son agrupados basándose en el principio de maximizar la similitud entre los elementos de un grupo, minimizando la similitud entre los distintos grupos, así se forman grupos tales que los objetos de un mismo grupo son muy similares entre si y al mismo tiempo son muy diferentes a los objetos de otro grupo.

Correlaciones: Son utilizadas para examinar el grado de similitud de los valores de dos variables numéricas, para ello se utiliza una fórmula estándar para medir la correlación lineal es el llamado coeficiente de correlación r, que es un valor real comprendido entre -1 y 1. Esto es que si r es 1 o -1 las variables están perfectamente correlacionadas, y si es 0 entonces no existe correlación alguna. En el caso de que r sea positivo, significará que la correlación existente entre dos variables refleja un comportamiento similar, ambas crecen o decrecen al mismo tiempo, y si es negativa, una variable crece y la otra decrece. Esta tarea es útil para establecer reglas de datos correlacionados.

Reglas de asociación. Consisten en crear en identificar relaciones no explicitas entre atributos categóricos, aunque hay muchas formas, la formulación más común es “si el atributo X toma el valor a, entonces el atributo Y toma el valor b”, Esta tarea no implica una relación causa-efecto, puede no existir una causa para que los datos estén asociados.

4.5 Tipos de Técnicas de Minería de datos

Para llevar a cabo las tareas descritas anteriormente existen diferentes tipos de técnicas de minería de datos.

Técnicas estadísticas y algebraicas: están basadas principalmente en expresar modelos y patrones mediante fórmulas algebraicas, funciones lineales, funciones no lineales, distribuciones, o valores agregados estadísticos como pueden ser varianzas, medias o correlaciones, entre otras. La mayor parte de las ocasiones estas técnicas cuando obtienen un patrón lo hacen mediante modelos ya predeterminado, de los cuales se estiman unos coeficientes o parámetros, por ello también son técnicas paramétricas. Dentro de los algoritmos más comunes de este grupo podemos mencionar a la regresión lineal –local o global-, la regresión logarítmica y la regresión logística. Los discriminantes lineales y no lineales basados en funciones predefinidas.

Técnicas Bayesianas: Se basan en estimar la probabilidad de pertenencia (a un grupo o clase), mediante la estimación de probabilidades condicionales inversas, utilizando el teorema de Bayes. Algunos algoritmos frecuentemente utilizados son el

clasificador bayesiano naive, los métodos basados en máxima verisimilitud y el algoritmo EM. Las redes bayesianas generalizan las topologías de las interacciones probabilísticas entre variables y permiten representar gráficamente dichas interacciones.

Técnicas basadas en conteos de frecuencias: se basan en contar la frecuencia en la que dos o más sucesos se presentan conjuntamente. Cuando un conjunto de sucesos es muy grande, entonces existen algoritmos que van separando grupos de pares de sucesos.

Técnicas basadas en árboles de decisión y sistemas de aprendizaje de reglas: son técnicas que además de su representación en formas de regla, se basan en dos tipos de algoritmos denominados "divide y vencerás", como el ID3/C4.5 o el CART, y los algoritmos, "separa y vencerás", como el CN2.

Técnicas relacionales, declarativas y estructurales: Este conjunto de técnicas representa los modelos mediante lenguajes declarativos, como los lenguajes lógicos, funcionales, o lógico-funcionales. Las técnicas de programación lógica inductiva y han creado el concepto de minería de datos relacional.

Técnicas basadas en redes neuronales artificiales: se trata de técnicas que aprenden un modelo mediante el entrenamiento de los pesos que conectan un conjunto de nodos o neuronas. La topología de la red y los pesos de las conexiones determinan el patrón aprendido.

Técnicas basadas en núcleo y máquinas vectoriales: una técnica que intenta maximizar el margen de los grupos o clases formadas mediante transformaciones que aumentan la dimensionalidad, llamadas kernels.

Técnicas escolásticas y difusas. La mayoría de las técnicas que junto a las redes neuronales forman la denominada computación flexible. Técnicas en las que los componentes aleatorios son fundamentales como los métodos evolutivos y genéticos o aquellos que utilizan funciones de pertenencia difusa.

Técnicas basadas en casos, en densidad o distancia. Son métodos que se basan en distancias al resto de los elementos, ya sea directamente, como los vecinos más próximos(los casos más similares) mediante la estimación de funciones de densidad. Entre los algoritmos más conocidos se encuentran el two-step o COBWEB y los no jerárquicos como K medias.

Todos los anteriores también forman una multitud de híbridos. En la tabla siguiente (tabla 4.1) se muestra la correspondencia entre técnicas y tareas de minería de

datos pudiendo apreciar que algunas tareas pueden ser resueltas por muy diversas técnicas, y algunas técnicas pueden aplicarse para varias tareas.

Nombre	PREDICTIVO		DESCRIPTIVO		
	Clasificación	Regresión	Agrupamiento	Reglas de Asociación	Correlaciones / Factorizaciones
Redes Neuronales	X	X	X		
Árboles de decisión ID3, C4.5, C5.0	X				
Árboles de decisión CART	X	X			
Otros árboles de decisión	X	X	X	X	
Redes de Kohonen			X		
Regresión lineal y logarítmica		X			X
Regresión logística	X			X	
Kmeans			X		
Apriori				X	
Naive BAyes	X				
Vecinos más próximos	X	X	X		
Análisis factorial					X
Twostep, cobweb			X		
Algorítmos genéticos y evolutivos	X	X	X	X	X
Máquinas de vectores soporte	X	X	X		
CN2 rules	X			X	
Análisis discriminante multivariante	X				

Tabla 4.1: Correspondencia entre técnicas y tareas de Minería de datos

4.6 Descripción de Técnicas de minería de datos

A continuación se describirán brevemente algunas técnicas de minería de datos, sólo aquellas que se pretenden combinar en conjunto con la prosopografía para logra el fin planteado por el presente trabajo.

La modelización estadística tiene por objetivo tratar de explicar el comportamiento de una variable en función del conocimiento de otras, dentro de este concepto se entiende que una variable tiene una cierta variabilidad y esta, se encuentra relacionada con el comportamiento de otras variables

Este modelo maneja dos tipos de variables: la variable de salida o también llamada de respuesta que es la variable que es centro de nuestro estudio y que generalmente se denota *y*, y las variables de entrada denominadas también explicativas, de las cuales mediante la función *r* depende el valor de nuestra variable de respuesta, denotadas por $x_1,x_2,\ldots x_j$.

$$y=r(x_1,x_2,\ldots x_j)$$

Esta técnica es la más utilizada en cuanto a la categoría estadística, ya que puede utilizarse tanto si el problema planteado consiste en predecir los valores de una cierta variable en función de algunos parámetros, como si el problema está en generar un modelo causal en el que las variables explicativas son la causa de la variación de la respuesta. Las siguientes variantes de esta técnica son las más frecuentemente utilizadas y que pueden ser de gran utilidad en la investigación histórica cómo más adelante lo veremos.

Regresión lineal. Trata de acercar los valores bivariantes de cierta modelización estadística a una función lineal que nos permita predecir o explicar el comportamiento de otros valores de la misma variable, mediante el cambio de sus parámetros, tiene por ventaja la posibilidad de trabajo sobre pocos datos, como desventaja es la poca flexibilidad a comportamientos complejos.

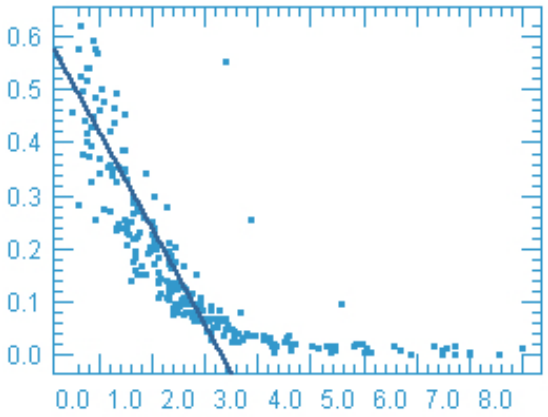


Figura 4.4: Ejemplo de Regresión Lineal

Discriminación lineal. Trata de analizar los valores bivariantes de alguna modelización estadística agrupándolos en áreas separadas por funciones matemáticas.

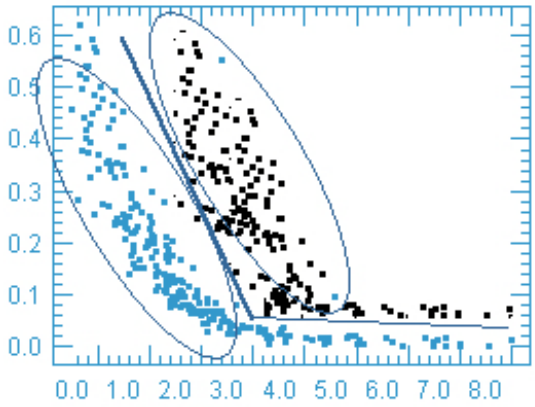


Figura 4.5: Ejemplo de discriminación Lineal

Regresión no paramétrica. Acerca los valores bivariantes de cierta modelización estadística a una función, no lineal flexible que se ajuste al comportamiento de los valores aún cuando estos sean complejos.

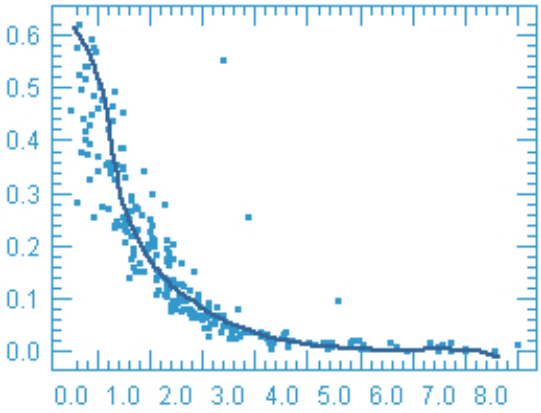


Figura 4.6: Ejemplo de regresión no paramétrica

Reglas de Asociación

Las reglas de asociación son utilizadas frecuentemente para expresar patrones de datos de una base de datos. Estos patrones pueden servir para conocer el comportamiento general del problema que genere la base de datos y de esta manera se tenga la mas información que permita asistir en la toma de decisiones.

Una regla de asociación es una proposición probabilística sobre la ocurrencia de ciertos estados en una base de datos. Entonces una regla de asociación puede detonarse del tipo SI a ENTONCES b donde a y b son dos conjuntos de ítems disjuntos.

Las reglas de asociación suelen trabajar con dos medidas para determinar su calidad y funcionamiento, la cobertura o soporte de una regla se define como el número de instancias que la regla es capaz de predecir correctamente, y la Confianza o precisión mide el porcentaje de veces que la regla se cumple cuando se puede aplicar.

Reglas de asociación multinivel

Debido a que en algunos casos los datos están muy dispersos, es decir, existe una gran cantidad de atributos comparados con al pequeña cantidad de items presentes en cada registro es bastante difícil encontrar relaciones de interés. Una forma de resolver este problema es agrupar atributos en categorías de esta manera el aprendizaje de reglas se basa en estas categorías y es más sencillo encontrar reglas con niveles adecuados de confianza o cobertura.

Las reglas multinivel son reglas de asociación que utilizan varios niveles de conceptos para expresar las relaciones. Para utilizar una regla multinivel además de los datos se debe proporcionar una jerarquía de conceptos que contenga un árbol de relaciones entre los atributos. Una jerarquía de conceptos define una secuencia de relaciones entre conceptos más específicos a conceptos más generales.

Reglas de asociación secuenciales

Este tipo de reglas expresa patrones de comportamiento secuenciales que se dan en instantes distintos, aunque cercanos, en el tiempo. El aprendizaje de reglas de asociación secuenciales se basa en encontrar las secuencias más comunes. Una Secuencia se define formalmente como una lista de conjuntos de ítems de un mismo cliente ordenada por el tiempo

4.6.1 Árboles de Decisión

Un árbol de decisión es un diagrama que representa en forma secuencial condiciones y acciones; muestra qué condiciones se consideran en primer lugar, posteriormente cuales en segundo lugar y así sucesivamente. Este método permite mostrar la relación que existe entre cada condición y el grupo de acciones permisibles asociado con ella.

Dada una base de datos se construyen diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.

Los árboles de decisión son normalmente construidos a partir de la descripción de la narrativa de un problema. Ellos proveen una visión gráfica de la toma de decisión necesaria, especifican las variables que son evaluadas, qué acciones deben ser tomadas y el orden en la cual la toma de decisión será efectuada. Cada vez que se ejecuta un árbol de decisión, solo un camino será seguido dependiendo del valor actual de la variable evaluada. Se recomienda el uso del árbol de decisión cuando el número de acciones es pequeño y no son posibles todas las combinaciones

Un árbol de decisión tiene unas entradas las cuales pueden ser un objeto o una situación descrita por medio de un conjunto de atributos y a partir de esto devuelve una respuesta la cual es una decisión que es tomada a partir de las entradas. Los valores que pueden tomar las entradas y las salidas pueden ser valores discretos o continuos. Se utilizan más los valores discretos por simplicidad, cuando se utilizan valores discretos en las funciones de una aplicación se denomina clasificación y cuando se utilizan los continuos se denomina regresión.

Un árbol de decisión ejecuta una prueba conforme este se recorre hacia sus 'hojas' para así alcanzar una decisión. Suele contener nodos internos, nodos de probabilidad, nodos hojas y arcos.

Nodo Interno. Contiene una evaluación sobre algún tipo de valor de una de los atributos de la base de datos.

Nodo de probabilidad. Indica que debe de ocurrir un evento aleatorio de acuerdo a la naturaleza del problema.

Nodo hoja. Representa el valor que devolverá el árbol de decisión.

Ramas. Plantean los posibles caminos que se tienen de acuerdo a la decisión tomada, justificando su recorrido tal decisión.

Pasos generales para la construcción de un árbol de decisión.

1. Seleccionar un conjunto de datos de la base de datos sobre el cual se aplicará el algoritmo de construcción del árbol de decisión.
2. Seleccionar el atributo objetivo, el atributo por el cual se clasificarán los casos de entrenamiento.
3. Descartar a priori los atributos irrelevantes para la clasificación.
4. Construir de manera recursiva el árbol de decisión.
    a) Si todos los casos de entrenamiento corresponden a objetos de una misma clase se ha logrado una buena clasificación. Se ha alcanzado una hoja del árbol de decisión.
    b) Si no encontramos un atributo por el que poder ramificar o se cumple alguna condición de parada, no se sigue expandiendo el árbol por la rama actual.
    c) Usando una tabla de casos se emplea alguna regla de división para seleccionar un atributo por el que ramificar. Para cada valor permitido de ese atributo obtenemos un subconjunto de casos en los que el atributo toma dicho valor y se genera un subárbol correspondiente recursivamente.
5. Poda a posteriori del árbol de decisión obtenido.
6. Generación de reglas a partir del árbol de decisión

Los pasos previamente descritos cuentan con variantes en cuanto a forma, y método a partir de diferentes algoritmos creados para la resolución de árboles de decisión, a continuación se mencionan algunos de estos algoritmos explicando brevemente el funcionamiento de dos de ellos –los más populares-

Algoritmo ID3

Este algoritmo es el método más famoso de todos los que existen para la creación de árboles de decisión. Usa una poda pesimista y utiliza el criterio de proporción de ganancia. Extensiones de ID3 le permiten tratar con datos erróneos e información incompleta.

Para que el árbol de decisión generado sea lo más sencillo posible, este algoritmo evalúa la capacidad de discriminación de cada uno de los atributos mediante el cálculo de las entropías de los distintos atributos en los casos de entrenamiento empleados. La entropía nos da una idea de la desorganización de la información, es decir, nos indica la capacidad de discriminación de cada atributo.

ID3 utiliza un método iterativo para construir árboles de decisión y prefiere los árboles sencillos frente a los más complejos (ya que, en principio, aquéllos que tienen sus caminos hasta las hojas más cortos son más útiles a la hora de clasificar entradas). En cada momento se ramifica por el atributo de menor entropía y el proceso se repite recursivamente sobre los subconjuntos de casos de entrenamiento correspondientes a cada valor del atributo por el que se ha ramificado.

Algoritmo C4

Está basado en ID3 y permite atributos continuos, sobre los que se aplican pruebas de la forma atributo<valor .Utiliza la poda pesimista

4.6.2 Minería de datos basada en grafos (Graph Mining)

Dada la necesidad de contar con una minería de datos estructurados, y teniendo a los grafos como una de las mejores estructuras de datos existente en el campo de las matemáticas discretas y las ciencias de la computación, la minería de datos basada en grafos se convierte en una muy importante herramienta.

El objetivo principal del GM como el de la minería de datos es extraer patrones de datos estructurados mediante el aprendizaje de subestructuras dentro de los grafos. Es decir la relación entre dos o mas entidades son vistas como grafos y mediante diversas aproximaciones se observan diferentes subgrafos y análisis de los mismos.

Las aproximaciones que se tienen de Graph Mining se encuentran:

- → Aproximaciones basadas en búsqueda Voraz
- → Aproximación basada en ILP (programación lógica Inductiva)
- → Aproximación basada en bases de datos inductivas
- → Aproximación Basada en Teoría de Grafos
- → Aproximación Basada en funciones núcleo

La minería de datos basada en grafos puede estar presente en diversas áreas de aplicación tales como “Cheminformatics”: Ingeniería del Software: Análisis de programas, Compuestos químicos, “Bioinformatics”: Estructuras proteínicas & bio-pathways, Análisis de redes de flujo (tráfico, workflow), Bases de datos semiestructuradas, p.ej. XML, Gestión del conocimiento: ontologías y redes semánticas CAD: Diseño de circuitos electrónicos (ICs), Sistemas de información geográfica y cartografía, Redes sociales. Estas últimas serán objeto de un caso de estudio particular al objetivo de este trabajo.

5 Caso de estudio Minable: Proyecto Autonomía

5.1 Planteamiento.

Se cuenta con un conjunto de datos, y fichas obtenidos a partir de la investigación del historiador, esta colección de datos carece de una clasificación sólida, estos datos son analizados por el investigador, y a través de sus conocimientos de prosopografía y su experiencia generan un conjunto de gráficos y tablas que sirven de base para la construcción de su contexto histórico y la justificación o refutación de la hipótesis de su trabajo de investigación.

Se plantea organizar los datos existentes y generar una metodología para la inserción de nuevos datos asimismo, previo estudio de la técnica de prosopografía y aprendizaje de diversos algoritmos de minería de datos se plantea ofrecer un conjunto de tablas, gráficos tendencias que representan el conocimiento obtenido a partir de la minería de datos.

5.2 Diseño Conceptual

5.2.1 Identificación de entidades.

Los datos almacenados para su estudio están basados en una lista en la cual un hombre, para el caso de estudio llamado 'Representante', tiene asignado un cargo en un año determinado, este representante puede figurar con otro cargo en años subsecuentes.

Uno de los objetivos del proyecto de investigación histórica es el de tener un seguimiento del historial de cargos que ha tenido un representante popular a lo largo de su carrera política así como detectar aquellos representantes que han estado en el panorama político por más tiempo, por lo que es preciso contar con una entidad llamada **Cargos** que nos enliste los representantes populares, y los cargos que tuvieron a lo largo del periodo de estudio.

Nuestro ente de estudio principal es el Representante, aquel sujeto histórico del cual el historiador ha logrado conjuntar una serie de datos de carácter personal, familiar, político y económico que le permita reconstruir su contexto y obtener conocimiento a través del manejo y análisis de dicha información y las relacionales posibles con otros Representantes.

Es necesario contar con una entidad que permita almacenar la información antes mencionada para cada uno de **Representantes** populares.

Dentro de la información trascendente para el historiador figuran los oficios o actividades que desempeñaban los representantes populares además de su cargo político ya que esto permite reconstruir relaciones y posturas.

Para su estudio estos **Oficios** y actividades se agrupan en categorías, aún cuando es importante conocer puntualmente el oficio desempeñado, lo es también conocer la categoría en la que el historiador lo ubica para así extraer información de carácter estadística, es por ello que esta se convierte en una entidad más que requiere incluir al representante popular, el oficio, y la categoría a la que se le asigna.

De manera paralela a los cargos un representante popular va cambiando su postura política a lo largo de su carrera, la información obtenida por el investigador histórico precisa ser almacenada de tal forma que se logre también un seguimiento de los cambios de **Posturas** de los representantes dado el año en el que se tuvo certeza de que el representante asumió tal o cual.

Una serie de datos muy importantes para la reconstrucción del contexto histórico es la definición de Personajes Relacionados con los sujetos de estudio, ya sean Representantes mismos u otros representantes Populares. Es preciso contar con una entidad que registre estas relaciones.

5.2.2 Identificación de atributos

Dado que se requiere hacer una biografía colectiva de los ciudadanos que tuvieron algún cargo político en el periodo de estudio, la prosopografía incluye la recolección de la información en cuatro niveles para la entidad RepresentantesP.

Personal
- Clave:
- Apellido Paterno
- Apellido Materno
- Nombre
- Año de Nacimiento
- Lugar de origen
- Patrono

Familiar
- Padre
- Madre
- Esposa

Económica
- Propiedades
- Albacea
- Herencia
- Heredero
- Bienes de la esposa
- Capital
- Biblioteca

Política
- Discursos
- Documentos
- Manifiestos

El manejo de la lista que compone los cargos de los representantes populares en sus respectivos años requiere de los siguientes atributos.

Entidad Cargos:

Clave
RepresentanteP
Año del Cargo
Entidad
Cargo

Con el fin de identificar plenamente los oficios y actividades que desarrollaban los representantes populares fuera de sus cargos se tiene:

Entidad Oficios:

Representante
Oficio
Categoría

Para el seguimiento del camino ideológico que se pudo rescatar de archivos y cuando se tiene la certeza de que un representante tomó alguna postura, es decir simpatizaba con tal o cual tendencia política dentro del contexto histórico, en un año determinado, es necesario considerar los siguientes atributos.

Entidad Posturas

Postura

Descripción

Para identificar las relaciones entre personajes los atributos son los siguientes.

Entidad PersonajesR
Clave
Personaje
Descripción

5.2.3 Identificación de Relaciones

SE_ENCARGA_DE: Cargos se relaciona con RepresentantesP ya que un representante puede ocupar varios cargos en años diferentes.

SE_DEDICA_A RepresentantesP está relacionado con oficios al haber varios posibles oficios que puede tener un representante y un oficio puede ser practicado por varios representantes.

TOMA: Posturas también se encuentra relacionada con RepresentantesP, debido a que se trata de llevar el seguimiento de las posturas de un representante popular, y una postura pudo ser tomada por diversos representantes en años diferentes.

CONOCE_A: PersonajesR está relacionada con RepresentantesP al vaciar en ella las relaciones de un representante con otro representante o persona fuera de la lista inicial pero que aún así merece ser registrada.

5.2.4 Identificación de Restricciones de Clave Primaria.

En la entidad Cargos un Representante popular puede aparecer varias veces con cargos diferentes o el mismo cargo, siempre y cuando el año del cargo no sea el mismo. Se le asigna un ID, a cada uno de los Cargos-entiéndase por Cargo a aquella tupla irrepetible que reúne un representante popular, un cargo y un año de cargo-.

La entidad RepresentantesP es considerada el eje central de toda la base de datos pues en ella transitan los datos de mayor relevancia y está relacionada con todas las demás entidades, cada Representante Popular es irrepetible y dado que en el universo particular de estudio existen homónimos es necesario incluir una clave única de identificación.

En la entidad Oficios un representante popular puede aparecer repetido varias veces siempre y cuando el oficio o actividad que realice sea diferente en cada tupla, por lo que aunque en función no se precisa de una llave primaria esta será un número consecutivo que identifique a un representante con su actividad.

La entidad posturas maneja una restricción de llave primaria similar a la anterior, es decir, un representante popular puede aparecer repetido varias veces siempre y cuando la postura tomada varíe, o el año varié ya que se trata de llevar un seguimiento de las posturas conocidas por año. Por ello se asigna un ID a cada una de las posturas que toma el representante popular por cada cambio de año.

En la entidad PersonajesR un representante popular es ligado a otro personaje, ya sea del universo de estudio central -RepresentantesP- o ya sea un personaje destacado del periodo y lugar de estudio definido, por lo que un representante popular puede repetirse siempre y cuando el personaje con el que está relacionado cambie. Se asigna un ID consecutivo para identificar tal relación.

5.2.5 Identificación de Restricciones de Cardinalidad.

Relación SE_ENCARGA DE:

RepresentanteP 1 SE_ENCARGA_DE N Cargos, un representante ocupa varios cargos en años diferentes.

Relación SE_DEDICA_A:

Representantep 1 SE_DEDICA_A N oficios, un Representante se dedica a varios oficios
Representantep N SE_DEDICA_A 1 oficios, un oficio puede ser desarrollado por varios RepresentantesP

Relación TOMA:

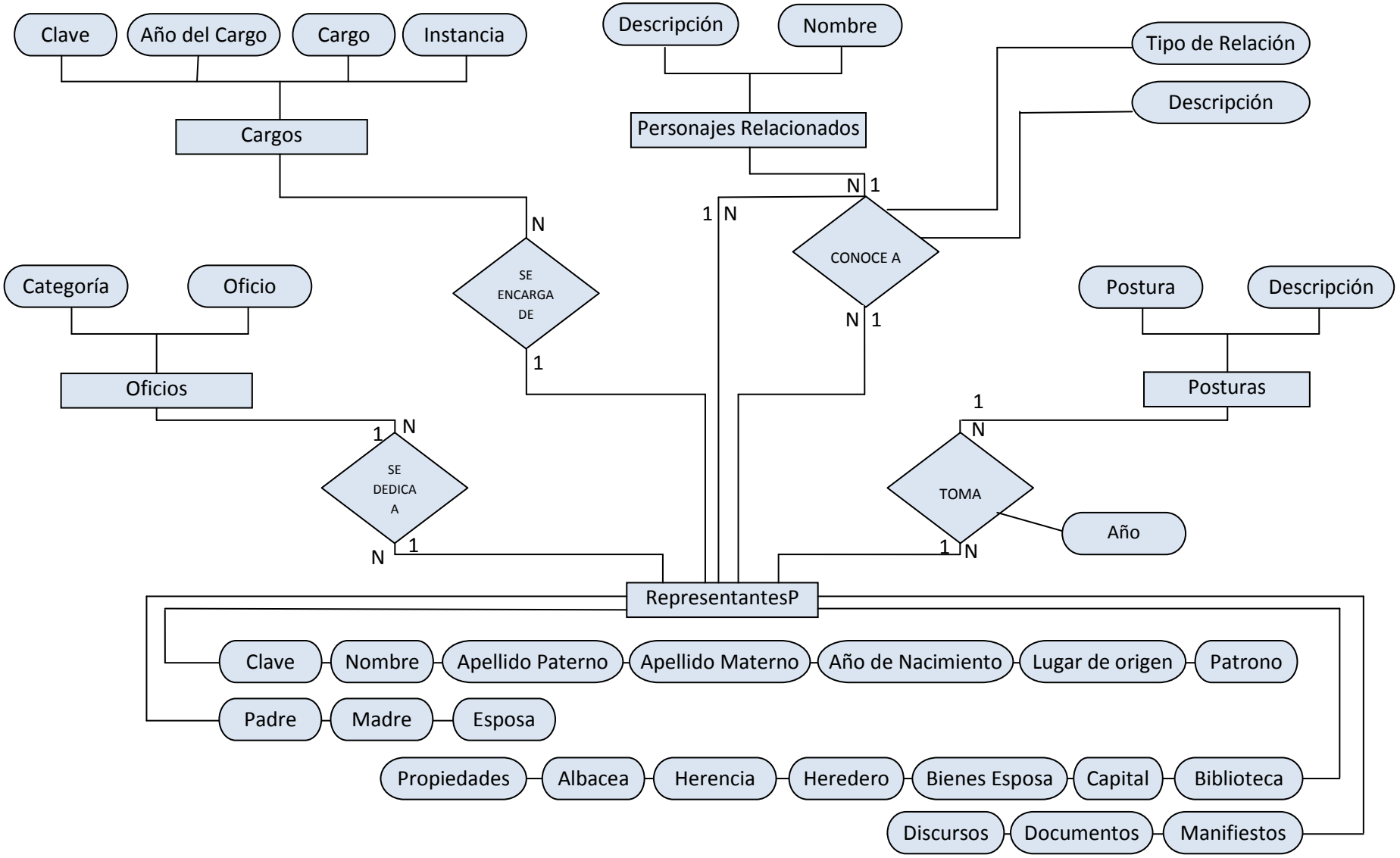
RepresentanteP 1 TOMA N Posturas, un representante toma varias posturas en años diferentes
RepresentanteP N TOMA 1 Posturas, Una postura puede ser tomada por varios RepresentantesP aún en el mismo año

Relación CONOCE_A

RepresentantesP. 1 CONOCE_A N RepresentantesP, un representante puede conocer a varios otros representantes
RepresentantesP. 1 CONOCE_A N PersonajesR, un representante puede conocer a varios Personajes

RepresentantesP. N CONOCE_A 1 RepresentantesP, Varios representantes conocen a un otro representante.
RepresentantesP. N CONOCE_A 1 PersonajesR, varios personajes conocen a un representante.

5.2.6 Diagrama Entidad-Relación.



5.3 Diseño Lógico

5.3.1 Traducción de tipos de entidades y relaciones.

Tablas procedentes de tipos de entidades del esquema E-R.

- **RepresentantesP** (*Clave, Apellido Paterno, Apellido Materno, Nombre, Año de Nacimiento, Lugar de origen, Patrono, Familiar, Padre, Madre, Esposa, Propiedades, Albacea, Herencia, Heredero, Bienes de la esposa, Capital, Biblioteca, Política, Discursos, Documentos, Manifiestos).

- **Cargos** (*Clave, Año del cargo, Cargo, Instancia).

- **Oficios** (*Clave, Oficio, Categoría).

- **Posturas** (*Clave, Postura, Descripción).

- **PersonajesR** (*Clave, Nombre, Descripción).

Tablas procedentes de tipos de relaciones del esquema E-R.

- **TOMA** (*id, RepresentanteP, Postura, Año).

- **CONOCE_A** (*id, RepresentanteP, PersonajeR, Tipo de Relación, Descripción).

- **CONOCE_A_REP** (*id, RepresentanteP, RepresentantePR, Tipo de Relación, Descripción).

5.3.2 Simplificación del esquema.

Observando como queda el esquema y dado que las relaciones generadas son de una a muchas, en el caso de la relación TOMA esta puede estar comprendida dentro de la entidad PersonajesR de tal manera que en ella misma se almacene la clave del representante popular y el año de la postura conocida, sin alterar la cardinalidad.

En la Relación CONOCE_A se puede aplicar un criterio de simplificación similar agregándole los atributos de la entidad PersonajesR , claro es para evitar duplicidad de los atributos Descripción, uno se refiere a la descripción del personaje, y el otro a la descripción de la relación. En tanto la relación

CONOCE_A_REP que concentra las relaciones entre elementos de la taba RepresentantesP únicamente incluirá la clave de cada uno de los elementos de las duplas relacionadas y será llamada simplemente Relaciones

Por lo tanto el esquema simplificado queda de la siguiente manera:

- **RepresentantesP** (*Clave, Apellido Paterno, Apellido Materno, Nombre, Año de Nacimiento, Lugar de origen, Patrono, Familiar, Padre, Madre, Esposa, Propiedades, Albacea, Herencia, Heredero, Bienes de la esposa, Capital, Biblioteca, Política, Discursos, Documentos, Manifiestos).

- **Cargos** (*Clave, Representantep, Año del cargo, Cargo, Instancia).

- **Oficios** (*Clave, RepresentanteP, Oficio, Categoría).

- **Posturas** (*Clave, RepresentanteP, Postura, Año Postura, Descripción).

- **PersonajesR** (*Clave, Nombre, Descripción, RepresentanteP, Tipo de Relación, Descripción).

- **Relaciones** (*Clave, RepresentantePA, RepresentantePB, Tipo de Relación, Descripción

5.3.3 Revisión de formas normales

Una tabla está en 1FN si sus atributos contienen valores atómicos. Al crear las tablas Relaciones, posturas y oficios se logró cumplir la condición de la 1FN ya que se garantizó que no existiera más de un valor por cada atributo.

Un esquema está en 2FN si: Está en 1FN.y todos sus atributos que no son de la clave principal tienen dependencia funcional completa respecto de todas las claves existentes en el esquema. En otras palabras, para determinar cada atributo no clave se necesita la clave primaria completa. La 2FN se aplica a las relaciones que tienen claves primarias compuestas por dos o más atributos. Si una relación está en 1FN y su clave primaria es simple (tiene un solo atributo), entonces también está en 2FN. Por tanto las tablas anteriores cuya clave primaria es simple cumplen con 2FN.

Para que una tabla esté en 3FN debe estar en 2FN y ningún atributo no-primario de la tabla es dependiente transitivamente de una clave candidata. Esto se garantiza ya que la información en todas las tablas es inherente a atributos muy específicos de los cuales no se puede tener dependencia transitiva, salvo en el caso de la información de

los representantes populares, la cual está perfectamente normalizada dentro de RepresentantesP

5.4 Diseño Físico.

En este apartado se muestra el diseño físico de la Base de datos Autonomía usando el SGBDR (sistema gestor de bases de datos relacionales) Access.

5.4.1 Breve resumen del SGBDR

Los tipos de campo, así como la definición de su tamaño (como se verá a continuación) permiten definir las restricciones de dominio que se refieren al tamaño y al tipo de los datos de un campo. Las reglas de validación ubicadas en las propiedades de los campos permiten especificar otras restricciones de dominio que limitan los valores del campo (por ejemplo, que no se admitan números negativos).

Para cada campo es posible especificar que no contenga valores nulos (es decir, imponer como restricción de dominio la eliminación del valor NULL del dominio del campo). También es posible especificar que si se trata de una cadena de caracteres, ésta no sea vacía.

Nombre de los campos

Deben estar identificados por nombres únicos dentro del contexto de la base de datos. Pueden tener hasta 64 caracteres con caracteres especiales y espacios en blanco (nunca al principio), pero no puntos, signos de exclamación o corchetes.

Tipos de campos

El tipo Texto permite datos de hasta 255 caracteres. El tipo Memo admite datos de hasta 65535 caracteres, pero tiene algunas limitaciones con respecto al anterior como, por ejemplo, que no se pueden crear claves sobre ellos. El tipo Numérico alberga datos numéricos tanto enteros como fraccionarios y números en coma flotante. El tipo de datos Fecha/Hora admite una representación conjunta para fechas y horas como un número fraccionario que el sistema interpreta con formato temporal y del que se puede extraer dicha información independientemente. El tipo Moneda se usa para datos relativos a cantidades monetarias. El tipo Autonumérico es un tipo que asigna automáticamente un valor único e identificador a los campos con este tipo (se usa a menudo para crear claves primarias). El tipo Sí/No es un tipo lógico con dos únicos posibles valores. El tipo Objeto OLE se usa para albergar imágenes, documentos y otros, con capacidad

hasta 1 GB. El tipo Hipervínculo se usa para albergar hipervínculos. Finalmente, el tipo Asistente para búsquedas permite definir los posibles valores que puede tener un campo de forma que el usuario pueda elegir valores de una lista predefinida.

Propiedades de los campos

Además del tipo de campo, es posible especificar otras propiedades de los campos (fig 5.1) , como su tamaño. Con el tamaño se consigue restringir aún más el tipo de campo para que concuerde con nuestras necesidades.

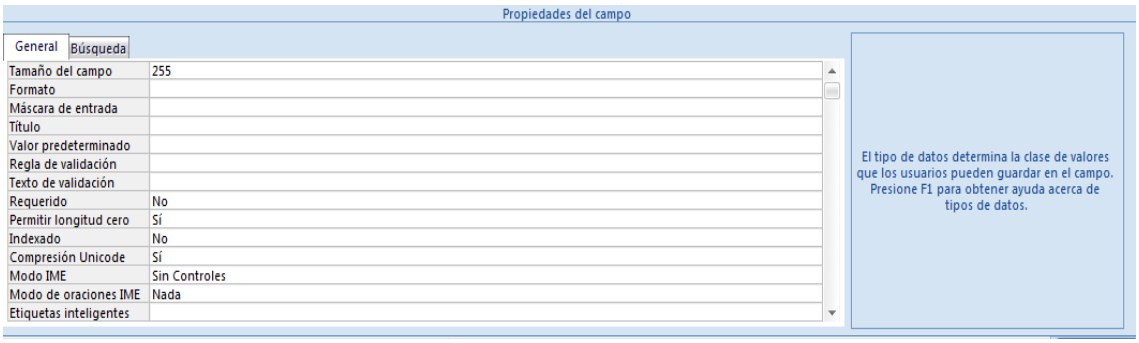


Figura 5.1: Propiedades de los campos en Access

Reglas de validación de los campos: asertos.

Las reglas de validación permiten especificar asertos que deben cumplirse para todos los valores de los campos. Estas reglas llevan asociado un texto de validación que permite informar al usuario del motivo por el que el contenido de un campo es incorrecto. Por ejemplo, la regla de validación puede ser >=0 (mayor o igual que cero.

Valores nulos.

La propiedad Requerido de un campo, si se establece a Sí, impide la inserción de valores nulos en los campos.

Índices.

Se pueden construir índices sobre campos aislados de una tabla o sobre un conjunto de ellos. Para construir un índice sobre un campo en concreto se indica en la propiedad Indexado el tipo de indexación que se desea. Se permiten índices con o sin valores duplicados. Un índice sin duplicados sobre un campo equivale a la especificación de una clave candidata.

5.4.2 Creación de las tablas de la Base de Datos Autonomía.

5.1 RepresentantesP		
Campo	Tipo	Tamaño
Clave*	Texto	6
Nombre	Texto	50
Paterno	Texto	50
Materno	Texto	50
AñoN	Número	--
LugarO	Texto	50
Patrono	Texto	50
Padre	Texto	100
Madre	Texto	100
Esposa	Texto	100
Propiedades	Texto	100
Albacea	Texto	100
Herencia	Texto	100
Heredero	Texto	100
BienesE	Texto	100
Capital	Texto	100
Biblioteca	Memo	--
Discursos	Memo	--
Documentos	Memo	--
Manifiestos	Memo	--

5.2 Cargos		
Campo	Tipo	Tamaño
Clave*	Texto	6
AñoC	Número	50
Cargo	Texto	50
Instancia	Texto	50
RepresentanteP	Texto	6

5.3 Oficios		
Id	Número	--
Oficio	Texto	50
Categorìa	Número	--
RepresentanteP	Texto	6

5.4 Posturas		
Clave	Número	--
RepresentanteP	Texto	6
Postura	Texto	50
AñoP	Número	--
Descripciòn	Memo	--

5.5 PersonajesR		
Clave	Número	--
Nombre	Texto	100
DescripciònP	Memo	--
RepresentanteP	Texto	6
TipoR	Texto	50
DescripciònR	Memo	--

5.6 Relaciones		
Clave	Número	--
RepresentantePA	Número	--
RepresentantePB	Número	--
TipoR	Texto	50
DescripciònR	Memo	--

Tabla 5.1 a 5.6: Diseño de las tablas de la BD

Figura 5.2 Diagrama de contexto nivel (0)

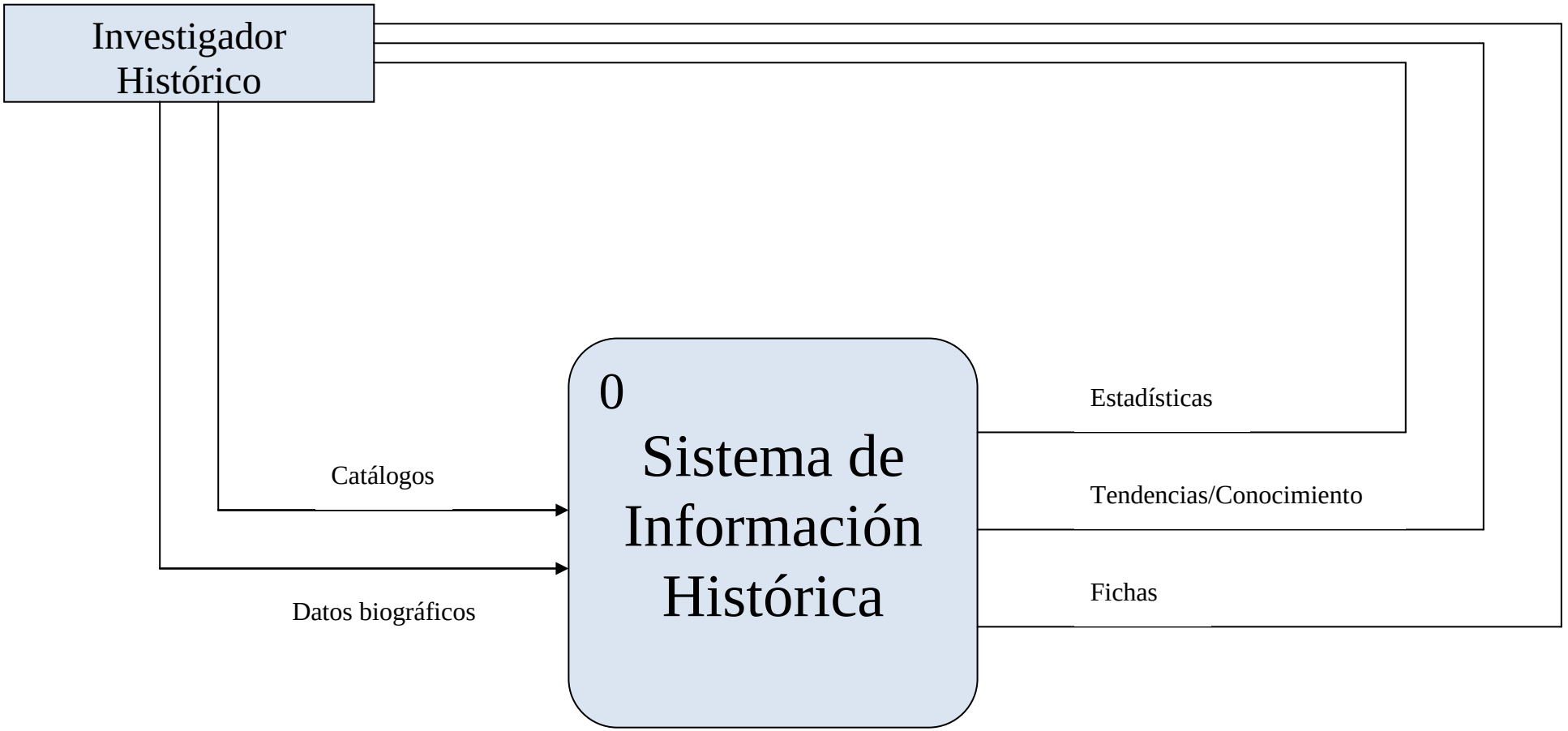


Figura 5.3 Diagrama de Análisis Nivel (1)

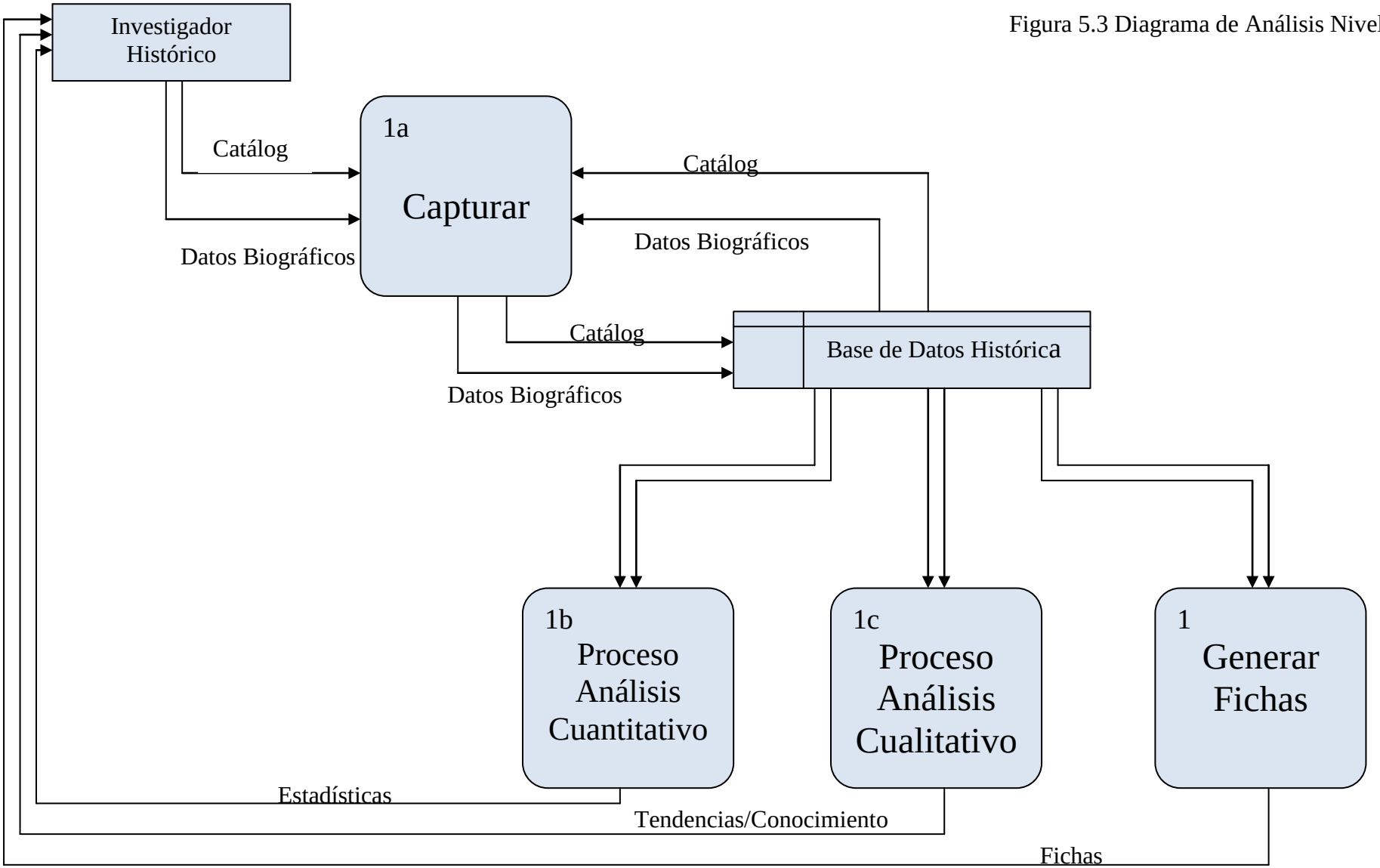
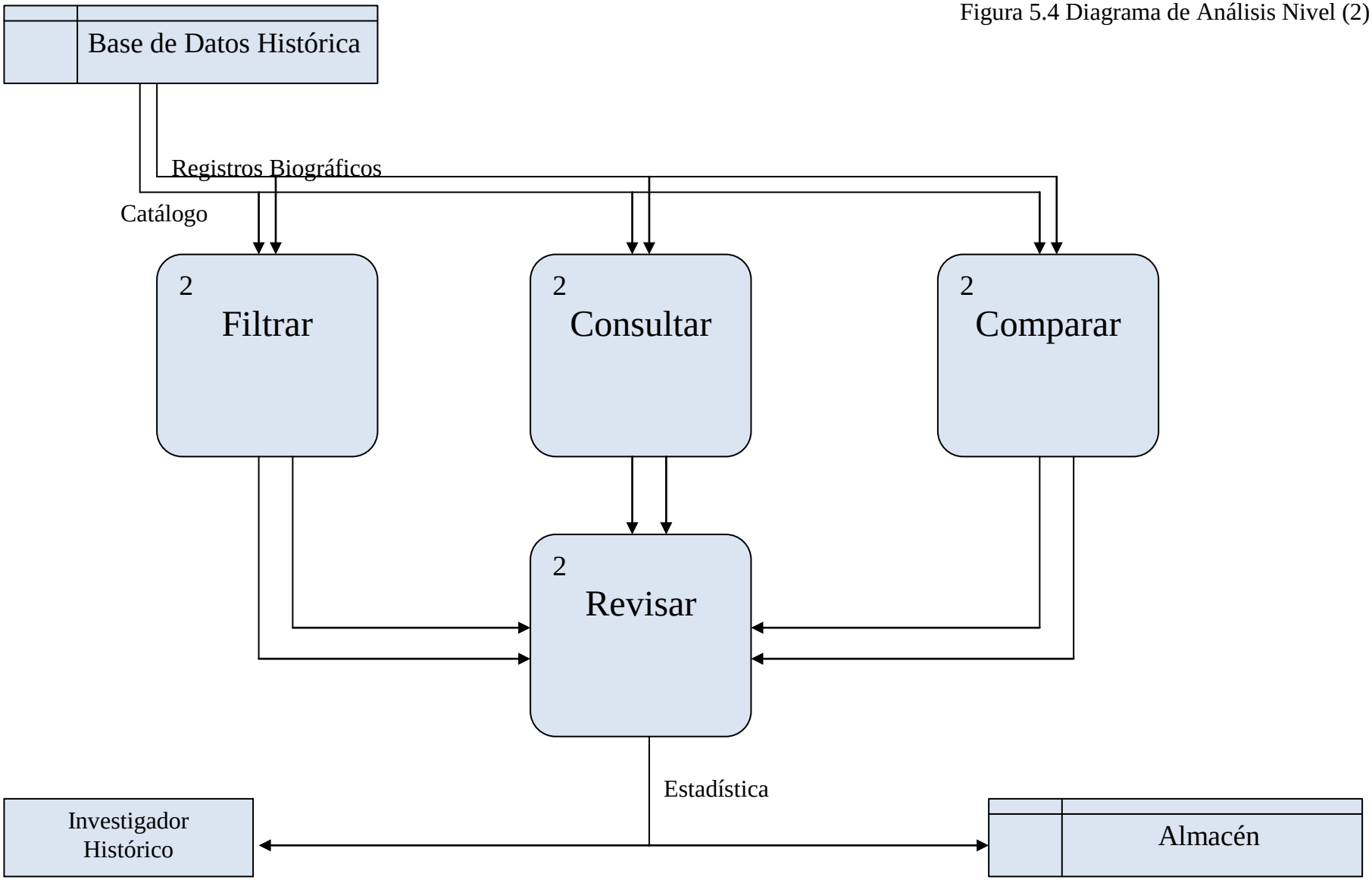


Figura 5.4 Diagrama de Análisis Nivel (2)



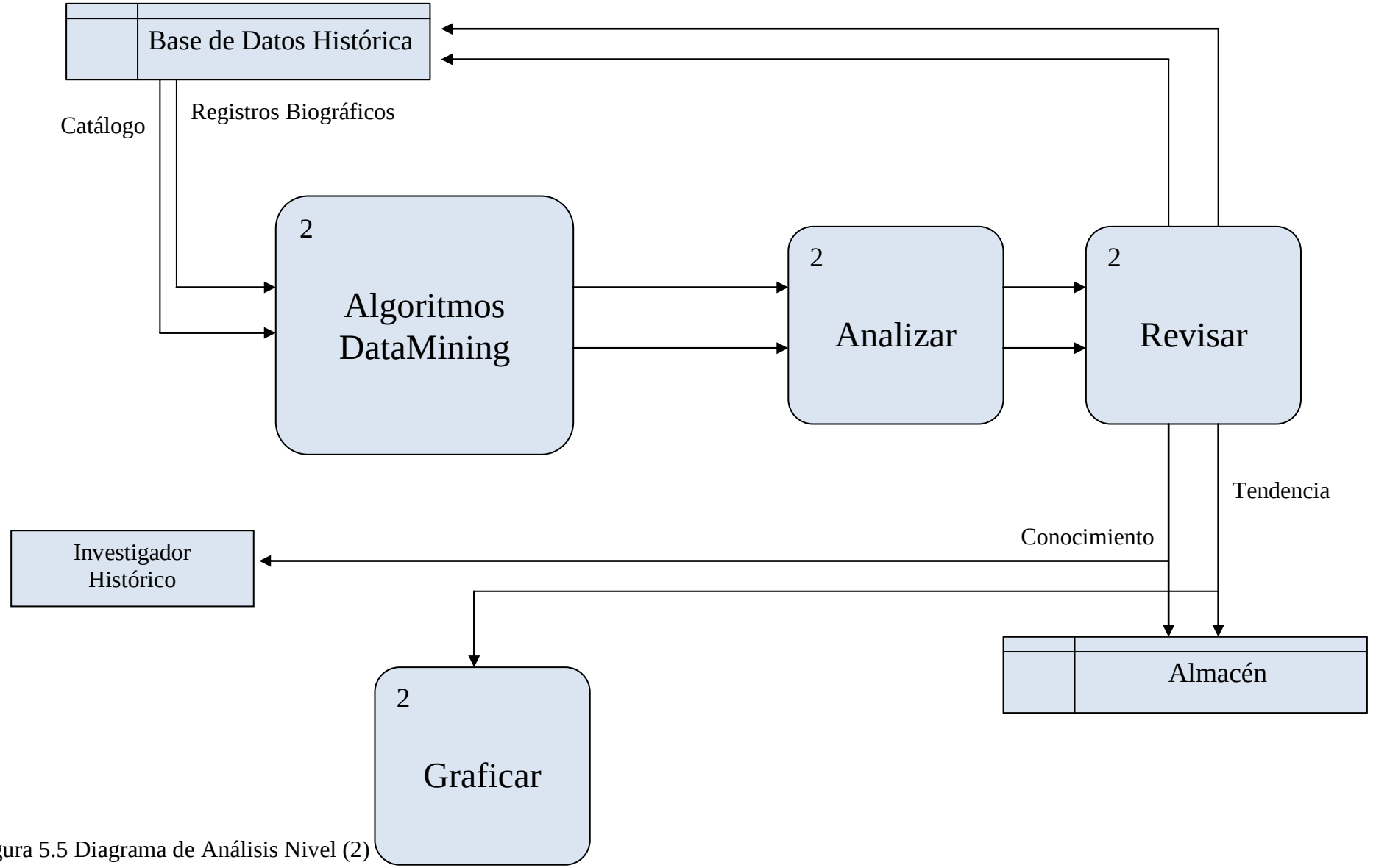


Figura 5.5 Diagrama de Análisis Nivel (2)

Figura 5.6 Diagrama de Análisis Nivel (2)

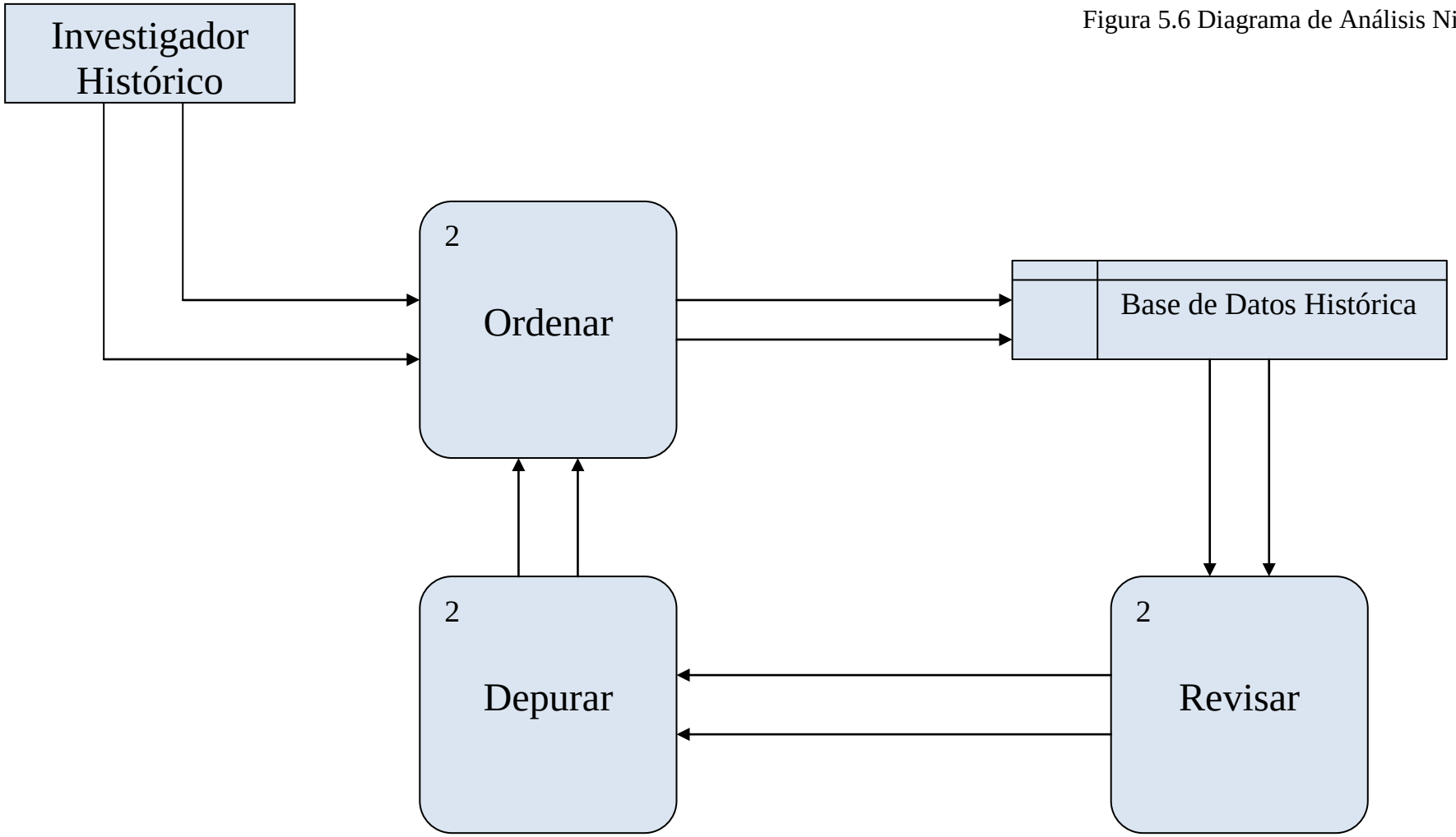
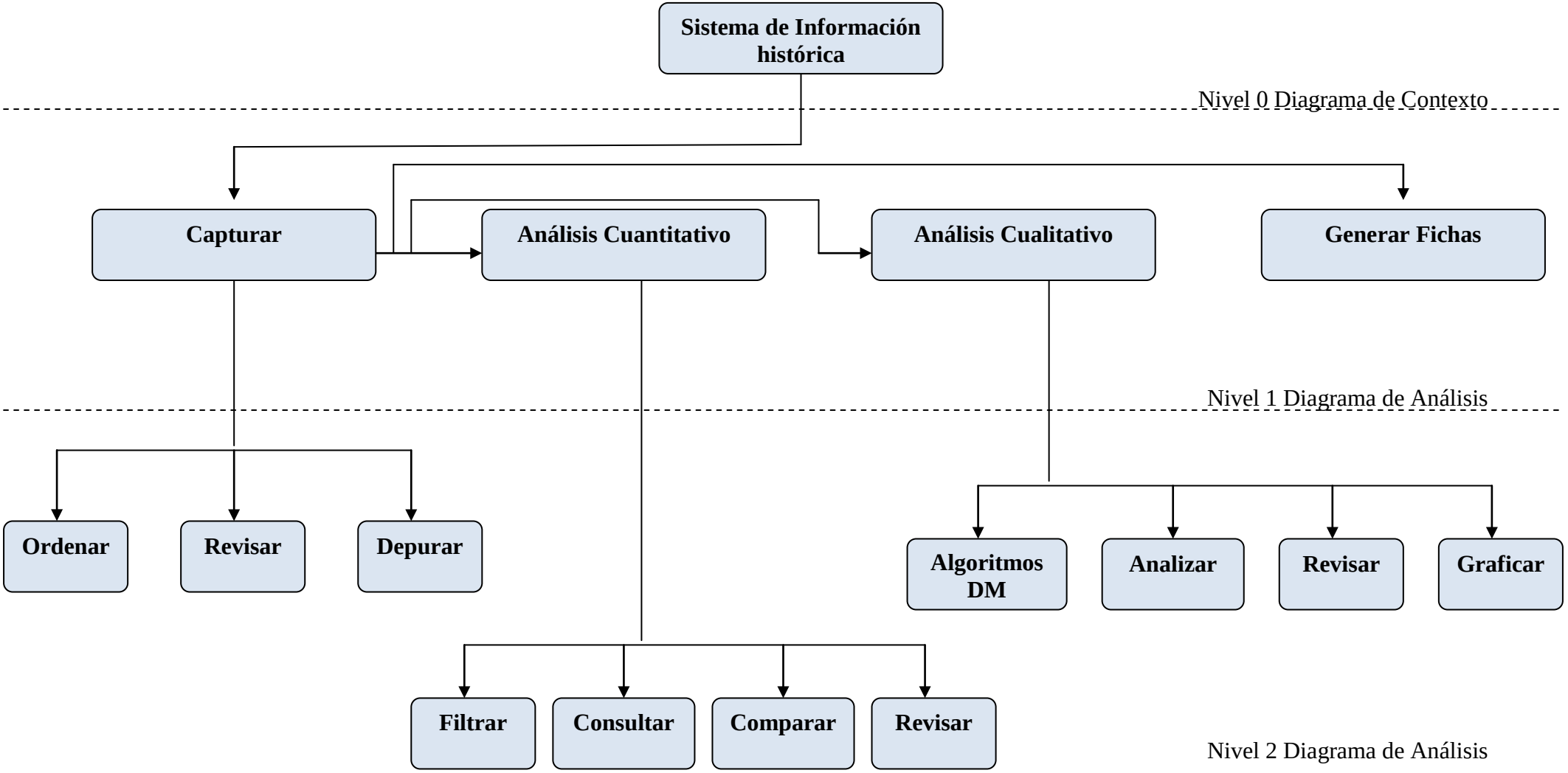


Figura 5.7 Diagrama de descomposición



5.6 Limpieza, transformación y preparación de los datos

Como se observó en el capítulo minería de datos, la fase de preparación de los datos es crucial y muy importante en el proceso de Minería de datos.

5.6.1 Integración y Normalización de Tablas

La base de datos Autonomía, pasó por un proceso de transformación a lo largo del presente trabajo para su normalización que permita.

- → Evitar la redundancia de los datos.
- → Evitar problemas de actualización de los datos en las tablas.
- → Proteger la integridad de los datos.
- → Aplicar los algoritmos de minería de datos
- → Analizar y generar reportes más confiables.

Base Autonomía Inicial: la base de datos autonomía diseñada inicialmente para la captura de los datos tenía la siguiente estructura:

Tabla 5.7 : Cargos

Atributos	Tipo
Paterno	Cadena
Materno	Cadena
Nombre	Cadena
Cargo	Cadena
Instancia	Cadena
Año	Entero

Tabla 5.8 : Cargos

Atributos	Tipo
Oficio	Cadena
Categoría	Entero

Tabla5.9 : RepresentantesP

Atributos	Tipo
Paterno	Nominal
Materno	Nominal
Nombre	Nominal
AñoC	
Cargo	
Instancia	
Oficio1	
Oficio2	
Oficio3	
Oficio4	
AñoN	Escala
Nativo	Nominal
Patrono	Nominal
Padre	Nominal
Madre	Nominal
Esposa	Nominal
Albacea	Nominal
Heredero	Nominal
CorrienteP	Nominal

Atributos	Tipo
BienesE	Nominal
Edad	Escala
Capital	Escala
Periodo	Discreto
Herencia	Nominal
RelacionesP	Nominal
RelacionesF	Nominal
RelacionesE	Nominal
Postura1	Nominal
Postura2	Nominal
Postura3	Nominal
Postura4	Nominal
Postura5	Nominal
Postura6	Nominal
Postura7	Nominal
Postura8	Nominal
Postura9	Nominal
Postura10	Nominal

Estas 3 tablas carecían de identificador único que hacía imposible el relacionarlas entre sí y optimizar el ordenado de los datos, la tabla RepresentantesP concentraba la mayor cantidad de datos y en el caso de Oficios y Posturas tenía una fuerte redundancia de datos, el caso de las relaciones es similar, la estructura definitiva quedó asentada en el capítulo anterior.

5.6.2 Reconocimiento

La primera acción a realizar una vez integrados todos los datos es un informe de estado de los atributos, por tabla, en este se muestran las características generales de los mismos.

Tabla 5.10: Reconocimiento de Cargos

Total de registros: 1017

Atributos	Tipo	#nulos	#dist.	Media	Desv. E.	Moda	Min	Max
Paterno	Nominal	0	295	-	-	García	Adorno	Zimbrelo
Materno	Nominal	691	113	-	-	Malpica	Aldana	Zubialdea
Nombre	Nominal	0	131	-	-	José María	Agustín	Vicente
Cargo	Nominal	0	8	-	-	Regidor	Alcalde	Síndico
Instancia	Nominal	0	7	-	-	Ayuntamiento	Ayuntamiento	Provinciales
Año	Escala	0	26	1824.5	7017	1834	1810	1835

Tabla 5.11: Reconocimiento de RepresentantesP

Total de registros: 440

Atributos	Tipo	#nulos	#dist.	Media	Desv. E.	Moda	Min	Max
Paterno	Nominal	0	295	-	-	García	Adorno	Zimbrelo
Materno	Nominal	308	113	-	-	Malpica	Aldana	Zubialdea
Nombre	Nominal	0	131	-	-	José María	Agustín	Vicente
AñoN	Escala	213	48	1781	119	1790	1750	1810
Nativo	Nominal	206	32	-	-	Puebla	Acatlán	Zacapoaxtla
Patrono	Nominal	398	10	-	-	Guadalupe	Guadalupe	Sra. Acatlán
Padre	Nominal	246	179	-	-	Diego Furlong	-	-
Madre	Nominal	248	178	-	-	Ana Malpica	-	-
Esposa	Nominal	266	172	-	-	Ana Hidalgo	-	-
Propiedades	Nominal	285	149	-	-	Tienda Mestiza	-	-
Albacea	Nominal	333	103	-	-	Joaquín de Haro	-	-
Heredero	Nominal	399	22	-	-	Hijos	-	-
CorrienteP	Nominal	235	13	-	-	Antiyorkino	-	-
BienesE	Nominal	351	30	-	-	Sin Dote	-	-
Edad	Escala	214	46	37	10.06	34	20	70
Capital	Escala	328	82	106656.6	174368	608000	300	608000
Periodo	Discreto	0	3	-	-	2	1	3

Tabla 5.12: Reconocimiento de Oficios

Total de registros: 825

Atributos	Tipo	#nulos	#dist.	Media	Desv. E.	Moda	Min	Max
Representante	Nominal	0	390	-	-	0	0	443
Oficio	Nominal	0	99	-	-	Comerciante	-	-
Categoría	Discreto	0	12	-	-	6	1	99

Tabla 5.13. Reconocimiento Posturas

Total de registros: 386

Atributos	Tipo	#nulos	#dist.	Media	Desv. E.	Moda	Min	Max
Representante	Nominal	0	239	-	-	312	0	441
Postura	Discreto	0	13	-	-	13	1	14
AñoP	Escala	0	15	-	-	1823	1812	1835

Tabla 5.14: Reconocimiento Relaciones

Total de registros: 495

Atributos	Tipo	#nulos	#dist.	Media	Desv. E.	Moda	Min	Max
RepresentanteA	Nominal	0	163	-	-	105	0	441
RepresentanteB	Nominal	0	160	-	-	420	0	441
TipoR	Discreto	0	79	-	-	F	F	P
DescripciónR	Nominal	86				amigo		

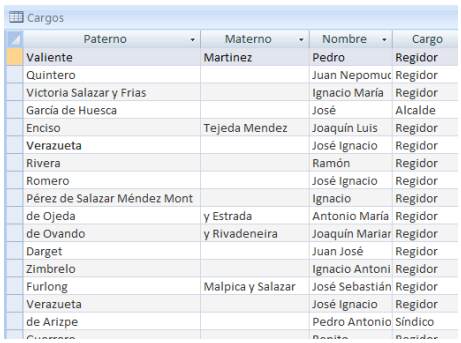
La tabla anterior es producto de una transformación de atributos necesaria para analizar las relaciones entre sujetos de la tabla Representantes, puesto que para el proceso de Graph mining que se aplicará es necesario contar con una matriz para el Graph matching. La tabla anterior era parte de la tabla RepresentantesP en la que además de existir redundancia de datos no habá la posibilidad de crear una matriz de relaciones mediante pesos de vectores.

5.6.3 Detección de valores faltantes

Como se puede observar en las tablas anteriores, el porcentaje de valores faltantes es grande, sobretodo en la tabla RelacionesP, es necesario encontrar una explicación o justificación de esta ausencia de datos para tomar políticas para su tratamiento:

En el caso particular de la investigación histórica, la gran diversidad de fuentes de información, y la complejidad en la búsqueda de datos particulares hace muy difícil contar con el 100% de los datos pertenecientes a un atributo, aún cuando podría

sugerirse entonces eliminar dicho atributo, los procesos y necesidades de información hacen preciso contar con dicho atributo mientras este cuente con el mínimo de datos que permitan al historiador reconocer o identificar un hecho histórico en particular.



Cargos

Paterno	Materno	Nombre	Cargo
Valiente	Martinez	Pedro	Regidor
Quintero		Juan Nepomuc	Regidor
Victoria Salazar y Frias		Ignacio María	Regidor
García de Huesca		José	Alcalde
Enciso	Tejeda Mendez	Joaquín Luis	Regidor
Verazueta		José Ignacio	Regidor
Rivera		Ramón	Regidor
Romero		José Ignacio	Regidor
Pérez de Salazar Méndez Mont		Ignacio	Regidor
de Ojeda	y Estrada	Antonio María	Regidor
de Ovando	y Rivadeneira	Joaquín Mariar	Regidor
Darget		Juan José	Regidor
Zimbrelo		Ignacio Antoni	Regidor
Furlong	Malpica y Salazar	José Sebastián	Regidor
Verazueta		José Ignacio	Regidor
de Arizpe		Pedro Antonio	Síndico

Figura 5.7: Muestra de datos dispersos

Apellidos Maternos

El 70% de los registros de apellido Materno es nulo (fig 5.7), este porcentaje es muy alto y podría sugerir eliminar el atributo o generar un atributo nuevo que conjunte el apellido paterno y el materno. Sin embargo un factor que debe analizarse muy en particular es la contextualización de los datos en la investigación histórica, es decir, no hay que olvidar que la base de datos es producto de la investigación de un universo de personas ubicado en un contexto temporal de principios del siglo XIX, donde la presencia de apellidos muy elaborados hace muy difícil de distinguir el apellido paterno del materno, tarea que queda a criterio del historiador, conocedor de las estructuras y formas de tales apellidos.

Datos de Nacimiento

El año y lugar de nacimiento, que tienen un porcentaje promedio de nulos del 40% de los datos son producto de una búsqueda en diferentes fuentes de información, actas, testamentos etc, que no necesariamente asentaban el lugar y año de nacimiento del Representante popular, sin embargo estos datos le darán una idea lo suficientemente acertada a los historiadores, mientras no esté por debajo del 50%.

Datos particulares (patrono, herencia, parientes, etc).

Estos datos aunque estén por debajo del 50% de aparición en la base de datos, son necesarios para determinar algunas preferencias de algunos de los representantes populares que permitan a los historiadores reconstruir su contexto histórico y económico.

5.6.4 **Detección de valores anómalos**

Un análisis de la tabla RepresentantesP comparada con los datos proporcionados por el historiador, nos arroja que pueden existir casos en los que la cantidad de datos

distintos esté fuera de la realidad o el contexto acotado. Este problema se aprecia muy claramente en el atributo CorrienteP, dado que la tabla de reconocimiento nos arroja 13 valores distintos, donde el universo debe estar acotado por sólo 6 posibilidades de corrientes., el siguiente histograma (fig.5.8) nos muestra claramente lo que está sucediendo. Valores que redundan a la misma idea pero con formatos diferentes o errores de captura.

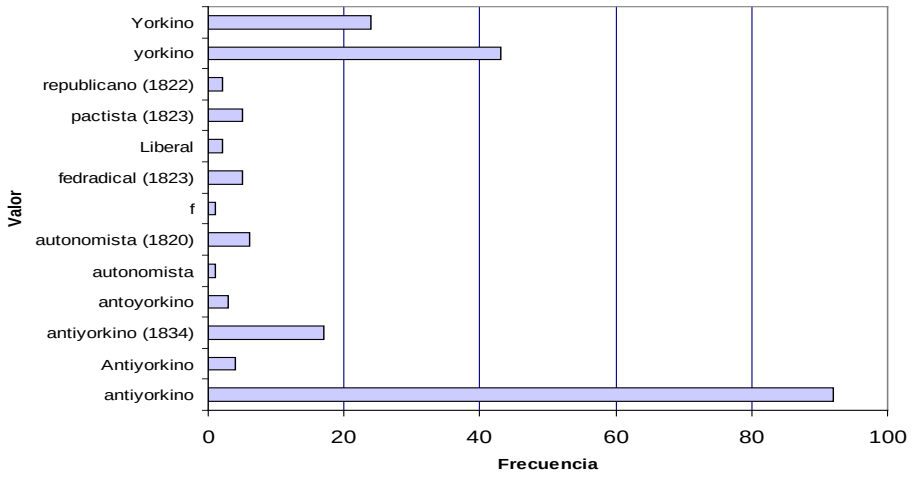


Figura 5.8: Histograma del atributo “Corriente”

El problema anterior fue detectado en muchos de los atributos en el interior de la base de datos, se deben solventar errores de captura, redundancias, abreviaturas, uso indiscriminado de mayúsculas y minúsculas, etc .

5.6.5 Discretización

Capital	BienesE
500,00 €	18 000 pesos
620,00 €	19 000 pesos
803,00 €	2 000 pesos
1.151,00 €	2 500 pesos
1.412,00 €	20 000 pesos
1.781,00 €	20 000 pesos
2.000,00 €	200 pesos
3.000,00 €	200 pesos
3.500,00 €	2000.00
3.530,00 €	26 000 pesos
5.000,00 €	26 000 pesos
5.000,00 €	3 000 pesos
5.000,00 €	3 000 pesos
5.700,00 €	300 pesos
6.000,00 €	300.00
6.000,00 €	4 000 pesos
6.955,00 €	4, 000 pesos

Figura 5.9: Datos Financieros de RepresentantesP

Los atributos referentes a datos financieros de los Representantes populares (fig 5.9) cuentan con una combinación de datos nominales y numéricos que nos impedirá establecer reglas de asociación y generar tendencias lineales sobre los datos y obtener modelos.

En el primer caso Capital, cuyos datos oscilan entre 300 y 608,000, la discretización permitirá generar reglas de asociación con otros atributos sin expandir la cantidad de posibilidades.

El primer paso para discretizar es definir un conjunto de rangos por el cual acotar los datos.

Rango	Valor
0-9999	Escaso
10000-99999	Medio
100000-596999	Considerable
597000-608000	Cuantioso

Tabla 5.15: Discretización de valores del atributo Capital

Este dato será sustituido en la tabla RepresentanteP mediante una sencilla instrucción SQL

```
UPDATE RepresentantesP Set Capital='Escaso' Where capital BETWEEN 0 and 9999

UPDATE RepresentantesP Set Capital='Medio' Where capital BETWEEN 10000 and 99999

UPDATE RepresentantesP Set Capital='Considerable' Where capital BETWEEN 100000 and 596999

UPDATE RepresentantesP Set Capital='Cuantioso' Where capital BETWEEN 597000 and 608000
```

Haremos lo mismo en el otro caso definimos nuestro rango y posteriormente ejecutamos una actualización de los datos. Para definir el rango y haer la selección de los valores que lo integrarán nos enfrentaremos a otro problema, la naturaleza inicial de este atributo no fue considerada de tipo numérico sino cadena que considera valores de la forma numérica '1000', '1000 pesos' 'sin dote' que hace difícil sustituirlos mediante una consulta simple por lo que es necesario primero normalizar los datos:

```
UPDATE Representantesp SET BienesE = REPLACE (BienesE,'pesos','')
```

Rango	Valor
Sin Dote.Huerfana	Sin Dote
0-999	Escaso
1000-4999	Medio
5000-50000,Mayorazgo	Cuantioso

Tabla 5.16: Discretización de valores del atributo BienesE

```
UPDATE RepresentantesP Set BienesE=’sin dote’ Where BienesE=’huerfana’

UPDATE RepresentantesP Set BienesE=’Escaso’ Where BienesE BETWEEN 0 and 999

UPDATE RepresentantesP Set BienesE=’Medio’ Where BienesE BETWEEN 1000 and 4999

UPDATE RepresentantesP Set BienesE=’50000’ Where BienesE =’mayorazgo’

UPDATE RepresentantesP Set BienesE=’Cuantioso’ Where capital BETWEEN 5000 and 50000
```

CAPITULO 6 IMPLEMENTACIÓN DE TÉCNICAS

6.1 Análisis de Cargos (Ordenamiento, selección y conteo)

A partir de la base de datos autonomía depurada y revisada se procede a obtener en primera instancia mediante técnicas básicas de conteo, ordenamiento, filtrado y estadística, un conjunto de tablas y gráficos que generen parte del conocimiento esperado.

Debido a que el año de 1820 fue coyuntural al ser un periodo de transición de poderes donde existieron dos ayuntamientos, las estadísticas deberán separarse en estos dos periodos, manejando resultados para el primer periodo en los años de 1810 a 1820, y para el segundo de 1820 a 1835, todo esto dentro de lo que comprende el análisis de los cargos. Cabe señalar que también deberá de incluirse como dato complementario una lista de representantes populares que ocuparon un cargo dentro del periodo 1786 a 1810, para generar únicamente una comparativa y un seguimiento de su influencia dentro de los años subsecuentes.

La primera necesidad dentro de la tabla Cargos está en obtener un historial de los cargos que ocupó cada uno de los representantes en el universo, esto nos permitirá conocer, los años en los que estuvo en el cargo, el cargo que tuvo, si repitió en años consecutivos, las personas o familias con mayor número de cargos durante el periodo de autonomía. Para ello mediante selecciones múltiples y ordenamientos se genera una variable temporal llamada Frec, variable entera que almacena el número subsecuente de cargo por cada año transcurrido, aumentando cuando el representante popular aparece de nuevo.

Cargos

Paterno	Materno	Nombre	Cargo	Instancia	Año	Clave	Frec
Aguilar		Ignacio	Diputado	Congreso Estatal	1834	001	0
Aguilar		Ignacio	Diputado	Congreso Estatal	1835	001	1
de Alducín		Juan Francisco	Regidor	Ayuntamiento	1820	002	0
de Alducín		Miguel	Regidor	Ayuntamiento	1813	003	0
Alfaro		José Mariano	Regidor	Ayuntamiento	1822	004	0
Alfaro		José Mariano	Regidor	Ayuntamiento	1823	004	1
Alfaro		José Mariano	Regidor	Ayuntamiento	1825	004	2
Altamirano		Mariano	Regidor	Ayuntamiento	1828	005	0
Altamirano		Mariano	Regidor	Ayuntamiento	1829	005	1
Altamirano		Mariano	Regidor	Ayuntamiento	1833	005	2
Altamirano		Mariano	Regidor	Ayuntamiento	1834	005	3

Tabla 6.1 Muestra de la Tabla Cargos ordenada por representante, con la variable Frec generada

Cargos							
Paterno	Materno	Nombre	Cargo	Instancia	Año	Clave	Frec
Azcarate		Juan Andrés	Síndico	Ayuntamiento	1810	023	0
Crespo	Sánchez de R..	Joaquín	Regidor	Ayuntamiento	1810	086	0
Darget		Juan José	Regidor	Ayuntamiento	1810	091	0
de Ovando	y Rivadeneira	Joaquín Marian	Regidor	Ayuntamiento	1810	127	0
Azcarate		Juan Andrés	Síndico	Ayuntamiento	1811	023	1
Crespo	Sánchez de R..	Joaquín	Regidor	Ayuntamiento	1811	086	1
de Arizpe		Pedro Antonio	Síndico	Ayuntamiento	1811	098	0
de Córdova	y Valdés	José Joaquín	Regidor	Ayuntamiento	1811	102	0

Tabla 6.2 Muestra de la Tabla Cargos ordenada por año, que muestra la generación de la variable temporal Frec

Gracias a la variable Frec también podemos conocer año por año el número de inclusiones de representantes nuevos, en los diferentes cargos posibles:

	1810	1811	1812	1813...
Total	22	23	20	30
Ingreso Nuevos	22	7	3	23
	1810	1811	1812	1813
Total Ayuntamiento	20	23	19	22
Ayuntamiento Nuevos	20	7	2	15
Ayuntamiento 1786-1810	18	15	12	12
	1810	1811	1812	1813
Total Cortes	2	0	1	8
Cortes Nuevos	2	0	1	8
	1810	1811	1812	1813
Total Congreso	0	0	0	0
Congreso Nuevos	0	0	0	0
	1810	1811	1812	1813
Total Provinciales	0	0	0	0
Provinciales Nuevos	0	0	0	0

Tabla 6.3 Muestra de los totales de conteo y ordenamiento de la tabla Cargos

6.2 Análisis de Oficios (Ordenamiento, selección y conteo)

La tabla Oficios, diseñada para registrar el historial ocupacional de cada representante, requiere un análisis muy particular para obtener el conocimiento que pueda ser útil para el historiador; en primera instancia se generó una variable que categorizó los oficios de acuerdo a una categoría planteada por el historiador de acuerdo a la relevancia y posibilidades políticas del oficio en cuestión.

Oficios		
RepresentanteP	**Oficio**	**Categoria**
0	Comerciante	6
1	Escribano	5
2	Subteniente	1
3	Eclesiástico	2
4	Capitán	1
5	Sastre	8
7	Abogado	3
8	Catedrático	4
9	Coronel	1
10	Abogado	3

Tabla 6.4: Muestra de Oficios con Atributo Categoría

El primer análisis incluye un agrupamiento y conteo sobre las categorías estos primeros se agrupan generando una consulta en SQL mediante sentencias simples tales como:

```
SELECT Categoria, COUNT(categoria) AS Frecuencia
FROM Oficios
GROUP BY Categoría;
```

Posteriormente se calculan porcentajes, porcentajes acumulados y totales, sobre el análisis de la variable categoría.

Categoría	Frecuencia	Porcentaje	Porcentaje acumulado
1	139	16,8	16,8
2	90	10,9	27,8
3	72	8,7	36,5
4	54	6,5	43,0
5	83	10,1	53,1
6	142	17,2	70,3
8	38	4,6	74,9
9	5	0,6	75,5
66	82	9,9	85,5
77	75	9,1	94,5
88	39	4,7	99,3
99	6	0,7	100,0
Total	825	100,0	

Tabla 6.5: Análisis Variable Categoría

De la misma forma se hace un estudio de frecuencias con las otras dos variables, en la primera *RepresentanteP*, es muy importante conocer estas frecuencias ya que para el historiador resulta interesante conocer a aquellos representantes con mayor número de oficios, en el caso de la segunda *Oficio*, solamente es necesario conocer los oficios con mayor número de incidencias.

Oficios

RepresentanteP	Oficio	Categoria
0	Capitán	1
0	Hacendado	66
0	Comerciante	6
0	Casateniente	88
1	Escribano	5
2	Subteniente	1
3	Eclesiástico	2
4	Casateniente	88
4	Labrador	66
4	Hacendado	66
4	Capitán	1

→

RepresentanteP	Frec	%	%val	%Ac
0	4	0,5	0,5	0,5
4	4	0,5	0,5	1,0
407	4	0,5	0,5	24,2
420	4	0,5	0,5	24,7
7	3	0,4	0,4	25,1
10	3	0,4	0,4	25,5

Tabla 6.6: Análisis de frecuencias variable Representante en Oficios

6.3 Análisis de Oficios (Reglas de Asociación)

Finalmente un reporte que resulta de mucha trascendencia en el análisis en los oficios de los Representantes populares es el de obtener las combinaciones más populares de categorías entre aquellos representantes con más de dos oficios.

Para ello utilizamos el algoritmo a priori para la obtención de reglas de asociación, como primer paso se filtran aquellos representantes cuya variable Frec (referente al número de oficios) es mayor o igual a 2 y se genera una tabla temporal en la que se incluye el detalle de las categorías en un arreglo para conocer las combinaciones más frecuentes., cabe señalar que existe la posibilidad de que un representante tenga dos oficios pertenecientes a la misma categoría por lo que esta podrá aparecer repetida, sin embargo se considerará también como una combinación. Por legibilidad las categorías son ordenadas en el arreglo de menor a mayor.

RepresentanteP	Categorías
0	1,6,66,88
4	1,66,66,88
5	8,99
7	1,3,66
8	4,88
9	1,8,8,88
10	3,88,88
11	3,6,8
12	66,88
13	1,3,66
	...

Tabla 6.7: Primer Paso A priori

Se obtiene un conjunto de datos relevantes L[1]con las categorías relevantes. A partir de L[1] se obtiene un conjunto de combinaciones candidatas C[2] con ítems relevantes.

L[1]
{6} [142]
{1}[139]
{2}[90]
{66}[83]
…

C[2]
{1,6}
{2,6}
{6,66}
{1,2}
{1,66}
{2,66}
…

Tabla 6.8: Segundo paso A priori

Se obtiene la relevancia de cada uno de los candidatos y se descartan aquellos con menos apariciones, generando un L [2], A partir de L[2] generamos un conjunto de candidatos C[3] con sólo aquellos elementos relevantes y sus combinaciones.

L[2]
{1,6}35
~~{2,6}3~~
{6,66}16
~~{1,2}4~~
{1,66}36
~~{2,66}6~~
…

C[3]
{1,6,77}(6)
{1,6,66}(3)
{1,66,77}(8)
{1,66,88}(4)
{1,3,66}(1)
{1,66,7}(2)
…

Tabla 6.9: Tercer paso A priori

Repetimos el proceso anterior para L[3] con C[3] y obtenemos finalmente C[4]=L[4]

C[4]
{77,1,66,77}(3)
{1,66,77,88}(1)
{1,66,88,06}(1)
{1,66,88,66}(1)
{1,66,88,77}(1)
{1,6,8,6}(1)
…

C[4]=L[4]
{77,1,66,77}(3)

Tabla 6.10: Cuarto paso A priori

Finalmente de cada uno de los pasos en L y ordenando las relevancias, el algoritmo a priori de Reglas de asociación nos arroja el siguiente resultado.

0166----	8
0601----	8
0106----	7
04------	6
08------	6
7706----	6
0305----	5
05------	5
0503----	5
0666----	5
0104----	4
0204----	4
020402--	4
0606----	4

0677----	4
77------	4
0108----	3
0202----	3
020202--	3
0266----	3
0404----	3
0505----	3
050505--	3
0688----	3
0806----	3
6601----	3
77016677	3
770601--	3

Tabla 6.11: Último paso A priori

6.4 Análisis de Posturas (Agrupamiento y selección, gráficas históricas)

La tabla posturas almacena las diferentes posturas políticas que el Representante fue tomando dentro del periodo de trabajo, incluyendo el año en el que se manifestó este cambio de postura. El objetivo del análisis de esta tabla consiste en conocer los años en los que más cambios de posturas hubo, así como aquellos representantes que más posturas tomaron.

Mediante agrupamiento y selección obtenemos aquellos Representantes con más cambios, y todos aquellos con más de una postura política se analizarán mediante una gráfica lineal grupal que nos permitirá conocer por observación, los puntos de transición.

Representante	Postura	AñoP
0	2	1823
0	1	1820
1	13	1835
2	1	1820
5	14	1829
5	12	1828
7	14	1833
7	14	1834
10	13	1834
11	13	1834
11	12	1828
13	13	1830

→

Representante	Postura	Porcentaje
312	8	2,1
186	6	1,6
268	6	1,6
104	5	1,3
212	5	1,3
20	4	1,0
83	4	1,0
95	4	1,0
142	4	1,0
267	4	1,0
360	4	1,0
407	4	1,0
435	4	1,0

Tabla 6.12: Selección de Representantes con más cambios de postura

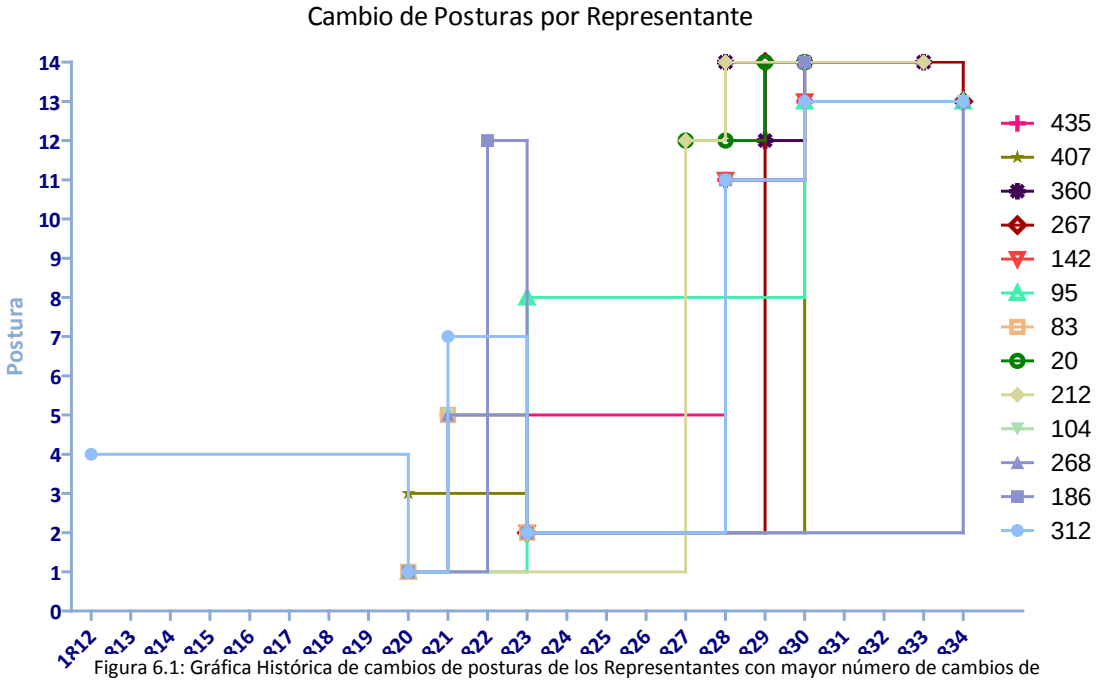


Figura 6.1: Gráfica Histórica de cambios de posturas de los Representantes con mayor número de cambios de postura

Para hacer un cruce de información y conocer los cambios de postura de una corriente política en particular, hacemos una consulta en la tabla Representantes.

Clave	CorrienteP
20	yorkino
83	yorkino
95	antiyorkino
104	Antiyorkino
142	Antiyorkino
186	Yorkino
212	Yorkino
267	Antiyorkino
268	antiyorkino
312	antiyorkino
360	Yorkino
407	Yorkino
435	antiyorkino

Tabla 6.13: Muestra de consulta de la corriente política de los representantes

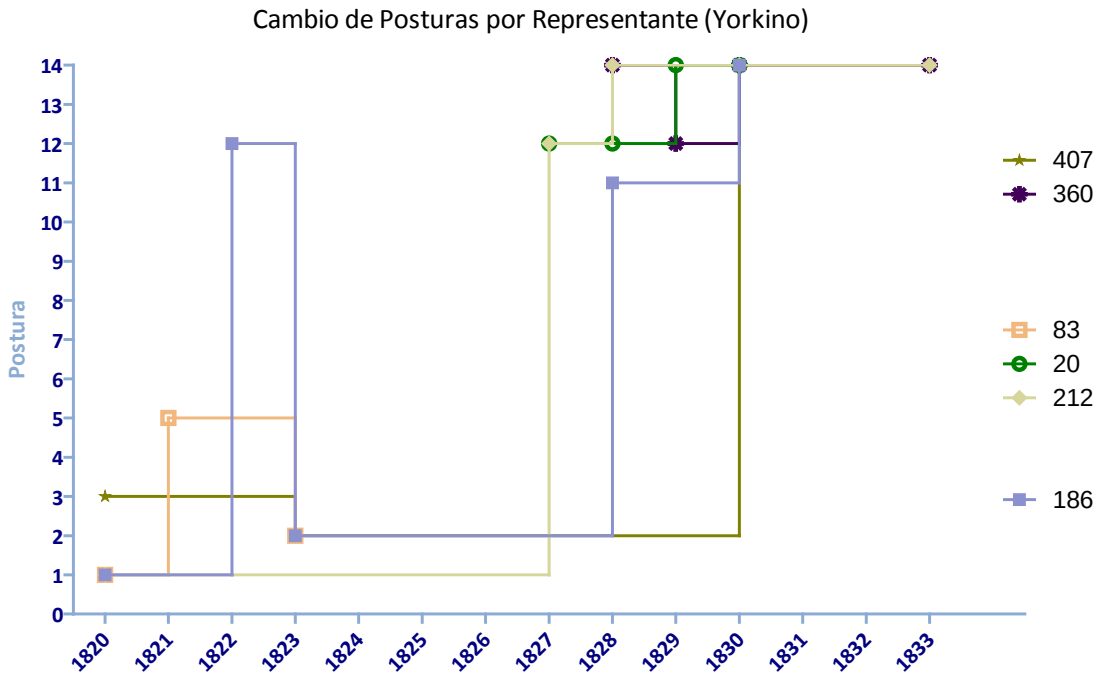


Figura 6.2: Gráfica Histórica de cambios de posturas de los Representantes Yorkinos

6.5 Análisis de Representantes (Redes bayesianas)

La relación existente entre la edad de los representantes, su capital, y la frecuencia de sus cargos en el periodo

Se tiene un campo denominado AñoN en el cual está determinado el año de nacimiento del representante popular, el campo Edad se calculó utilizando el primer año en el que el representante ocupó un cargo. Trataremos de buscar una relación existente entre su edad, su capital, el número de cargos y la CorrienteP a la que perteneció, y trataremos de averiguar si las primeras son consecuencia de la última o viceversa. Para ello utilizaremos de manera muy sencilla la teoría de redes bayesianas auxiliados con la utilidad online D-trail de B-Course.

Una red Bayesiana es una tupla B=(G, Θ). Donde G es el grafo y Θes el conjunto de distribuciones de probabilidad P(X|Pa(X)) para cada variable desde i=1 hasta n y Pa(X) representa los padres de la variable X en el grafo G,.

Primero seleccionamos y separamos solo aquellas variables donde se garantice un reducido número de datos nulos. (Tabla 6.14)

Clave	Nativo	Corriente	Edad	Capital	Cargos
11	Puebla	antiyorkino	33	24000	3
69	Puebla	antiyorkino	29	36000	3
180	Apetatiltla	antiyorkino	37	76000	5
232	Tlaxcala	Antiyorkino	31	17000	7
267	Puebla	antiyorkino	33	11000	3
287	Amozoc	antiyorkino	27	5000	9
296	Puebla	antiyorkino	25	608000	1
334	Puebla	antiyorkino	40	50000	3
375	Puebla	antiyorkino	50	19000	2
435	Puebla	antiyorkino	40	19842	6
179	Veracruz	antiyorkino	27	173000	4
337	Puebla	antiyorkino	27	608000	6
373	Puebla	antiyorkino	42	28000	2
206	Puebla	autonomista	45	10000	2
5	Puebla	yorkino	61	300	4
7	Tlaxcala	yorkino	27	17151	2
22	Puebla	yorkino	55	6000	1
72	Puebla	yorkino	28	15000	2
212	Veracruz	yorkino	40	9500	2
220	México	yorkino	46	8995	2
244	Puebla	yorkino	39	3000	6
293	Veracruz	yorkino	49	500	2
371	Puebla	yorkino	40	55000	4
		...			

Tabla 6.14: Selección de variables con menos número de datos nulos

Esta consulta es transformada en formato texto para introducirla en la herramienta D.Trail mediante su página de carga de archivo. Se descarta el campo clave, para que el análisis y estudio del conjunto de distribuciones de probabilidad no se confunda.

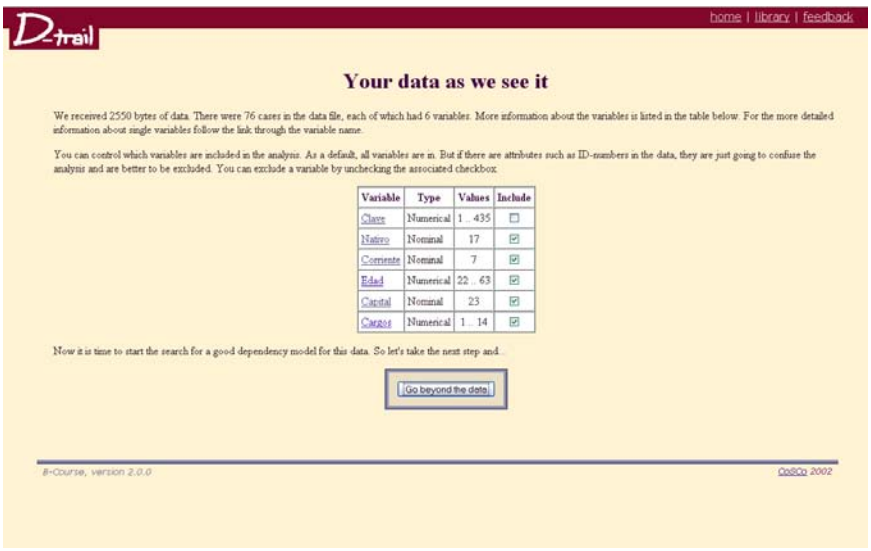


Figura 6.3: Carga de datos en el sistema D-Trail para elaboración de redes bayesianas

El sistema evalúa los posibles modelos candidatos de acuerdo a las distribuciones de probabilidad para cada variable, basados en la teoría de la probabilidad.

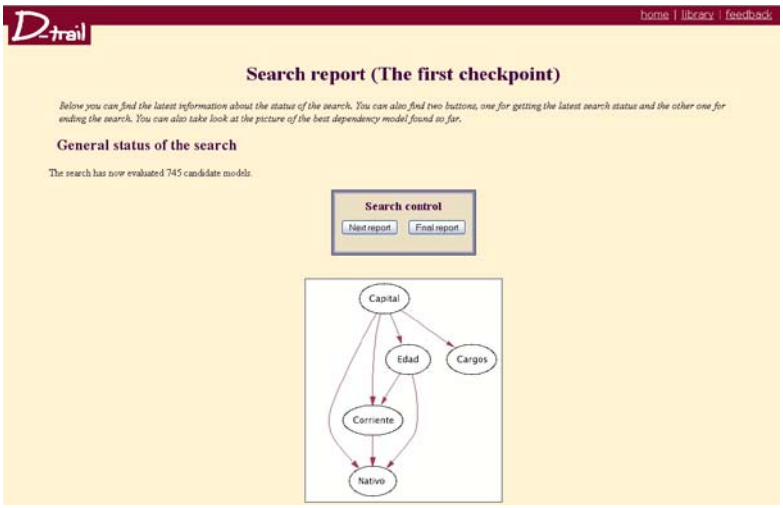


Figura 6.4: Evaluación de datos por D-Trail arrojando posibilidades de modelos de dependencia

Se revisan exhaustivamente todas las posibilidades de modelos de dependencia, quedando como resultado final el más óptimo.

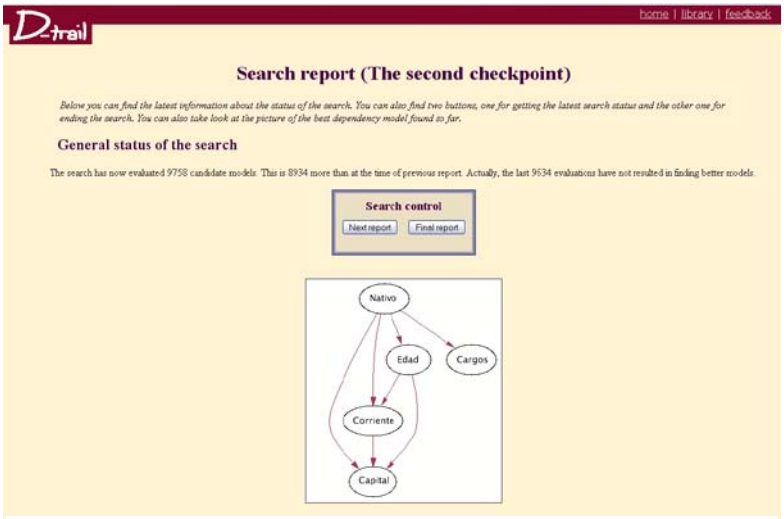


Figura 6.5: Selección del modelo de dependencia más óptimo

Una vez que se eligió el modelo se hacen procesos de inferencia de acuerdo a los porcentajes conocidos.

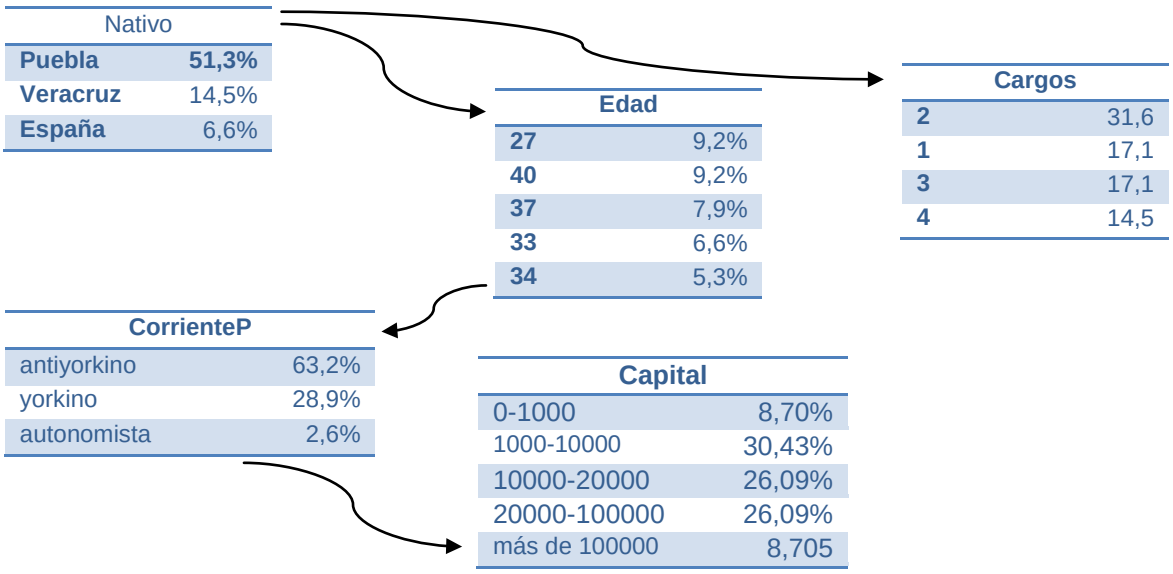


Figura 6.6: Procesos de inferencia

El análisis anterior nos presenta una serie de relaciones muy interesantes, nos permite demostrar que existe una relación directa entre el lugar de origen y la permanencia o reiteración de los cargos de un representante popular, ya que la mayoría de representantes con más cargos son originarios de Puebla. Por otro lado la corriente política parece tener una dependencia en la edad, por ejemplo en el caso de los antiyorkinos donde la edad promedio está entre 32 y 35 años. Siendo en esta corriente política donde más se concentra el capital.

6.6 Análisis de Relaciones (Conteo y selección)

Las relaciones existentes entre los representantes populares y su tipo son uno de los elementos más significativos dentro del análisis de los datos, ya que con el estudio de las mismas el historiador puede reconstruir el contexto político y plantearse la construcción de familias políticas y relaciones de interés.

Para ello dentro de la base de datos autonomía, se tiene la tabla Relaciones en la que se tiene una relación de la forma CONOCE(RepA,RepB,Tipo) es decir RepA conoce a RepB de forma Tipo[Familiar,Económica,Política].

Relaciones

Id	RepresentantePA	RepresentantePB	TipoR	DescripciónR
420	187	363	E	socio
421	263	367	F	suegro
422	369	443	F	amigo
423	61	371	E	socio fábrica
424	337	373	F	tutor
425	163	376	P	junta municipal

Tabla 6.15: Muestra Tabla Representantes

El primer objetivo es detectar aquellos Representantes que puedan ser considerados más influyentes, o con mayor cantidad de relaciones ya que ellos pueden considerarse líderes o principales organizadores de movimientos políticos autónomos. Mediante agrupamiento y selección conoceremos estos datos de frecuencia.

339	21
420	14
95	12
105	12
225	12
407	12
185	11
337	10
345	8
363	8

Tabla 6.16: Muestra de selección de representantes con más relaciones

Al tener esta información, esta se puede cruzar para generar parciales y de esta manera identificar aquellos Representantes cuya influencia sea sólo política, o solo económica. Seleccionando sólo aquellos con cierto tipo de relación.

Frec.	Tipo Rel.	Política
185	E	10
339	E	8
420	E	6
295	E	5
312	E	5
186	E	4

Tabla 6.17: Muestra de selección de representantes con más relaciones de tipo

6.7 Análisis de Relaciones (Similitudes vectoriales)

Para poder descubrir los grupos y redes políticas más importantes dentro del universo de los Representantes populares de este contexto se propone analizar mediante la medición de similitudes en modelos vectoriales esto significaría que cada representante significaría un vector de pesos no binarios de acuerdo al tipo de relación con los otros representantes. Se le asignará un peso constante a los tipos de relación en función de su trascendencia, por lo que la más importante Político tendrá un peso específico de 3, Económico un peso específico de 2 y Familiar un peso específico de 1 y en caso de no existir relación alguna se asignará un peso específico de 0. La tabla Relaciones se transformará en una matriz dispersa de pesos específicos.

	RepA(1)	RepA(2)	RepA(3)	RepA(3)	RepA(5)	RepA(6)	RepA(7)	…	RepA(n)
RepB(1)	-								
RepB(2)	[0,1,2,3]	-							
RepB(3)			-						
RepB(4)				-					
RepB(5)					-				
RepB(6)						-			
RepB(7)							-		
…								-	
RepB(n)									-

Tabla 6.18: Muestra de matriz dispersa de pesos específicos

La idea básica de este modelo de recuperación vectorial reside en la construcción de una matriz (podría llamarse tabla) de Relaciones, donde las filas las columnas correspondieran a los Representantes de todo el universo de estudio Así, las filas de esta matriz (que en términos algebraicos se denominan vectores) seran equivalentes al campo RepA de esta manera un Representante podría expresarse de la

manera:

· RepB(1)=(1, 2, 0, 0, 0, ..., 1, 3) : Siendo cada uno de estos valores el peso específico correspondiente a la relación con RepA.

La longitud del vector de RepB sería igual al total de términos de la matriz (el número de columnas), que en este caso sería el mismo que el número de vectores RepB.

La segunda idea asociada a este modelo es calcular la similitud entre cada uno de los vectores de tal forma que existiría un Factorial de n operaciones de comparación de vectores para generar aquellos con más similitudes.

Se dispone de varias fórmulas que nos permiten realizar este cálculo, la más conocida es la Función del Coseno, que equivale a calcular el producto escalar de dos vectores (A y B) y dividirlo por la raíz cuadrada del sumatorio de los componentes del vector A multiplicada por la raíz cuadrada del sumatorio de los componentes del vector B.

De esta manera se calcula este valor de similitud. Como es obvio, si no hay coincidencia alguna entre los componentes, la similitud de los vectores será cero ya que el producto escalar será cero (circunstancia muy frecuente en la realidad ya que los vectores llegan a tener miles de componentes y se da el caso de la no coincidencia con mayor frecuencia de lo que cabría pensar).

También es lógico imaginar que la similitud máxima sólo se da cuando todos los componentes de los vectores son iguales, en este caso la función del coseno obtiene su máximo valor, la unidad.

$$\frac{\sum_{i=1}^{n} Xi * Yi}{\sqrt{\sum_{i=1}^{n} Xi^2 * \sum_{i=1}^{n} Yi^2}}$$

La tabla resultante se ordena por similitudes detectando un grupo de personas con presencia e influencia en las demás.

Relación	Similitud
RepA(1)RepB(2)	S(1,2)
RepA(1)RepB(3)	S(1,2)
RepA(2)RepB(1)	S(1,2)
RepA(2)RepB(3)	S(1,2)
RepA(n)RepB(1)	S(1,2)
…	
RepA(n)RepB(n)	S(1,2)

Tabla 6.19: Tabla ordenada por similitudes

6.8 Análisis de Relaciones (Minería de grafos)

El siguiente objetivo dentro de la tabla relaciones será el generar una representación gráfica de la red social integrada por la tabla relaciones. Los analistas de redes sociales utilizan dos tipos de herramientas matemáticas para representar información sobre los patrones de relaciones entre actores sociales: grafos y matrices. Dado que el primer análisis lo utilizamos con matrices, y que al usuario final, el historiador, le resultará muy útil el observar los resultados en diagramas comprensibles nos enfocaremos al uso de grafos.

Los analistas de redes utilizan principalmente un tipo de representación gráfica que consiste en puntos (o nodos) para representar actores y líneas (o flechas) para representar lazos o relaciones. Cuando los sociólogos tomaron esta forma de representación de los matemáticos, renombraron sus gráficos como “sociogramas”.
Los matemáticos distinguen los diferentes tipos de representaciones gráficas con los nombres de “grafos recíprocos”, “grafos orientados” o simplemente “grafos”.

Existen muchas variaciones en los sociogramas, pero todos ellos comparten la característica común del uso de un círculo etiquetado para cada actor en la población que describimos y segmentos de línea entre pares de actores para representar el hecho que existe un vínculo entre ellos.

A partir de la tabla Relaciones armaremos entonces una red de grafos ponderados por el tipo de relación.

RepresentantePA	RepresentantePB	TipoR
163	203	P
163	281	E
163	318	E
163	376	P
218	318	F

Tabla 6.20: Muestra de Relaciones

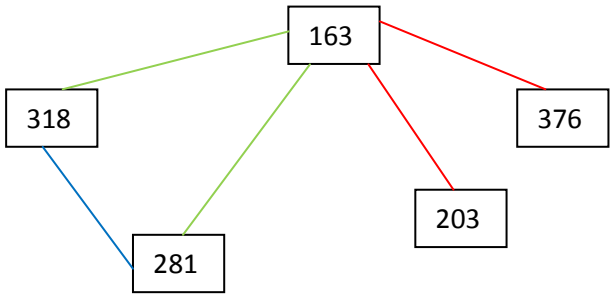


Figura 6.7: Construcción básica de los grafos de relaciones, donde las líneas de colores representan el tipo de relación ente cada nodo o representante

De esta forma Los nodos y las relaciones tendrán atributos propios permitiéndonos establecer criterios de selección, y representación de las relaciones. Con la ayuda del software de visualización de redes NetDraw obtenemos la red social generada con la totalidad de registros de la base de datos.

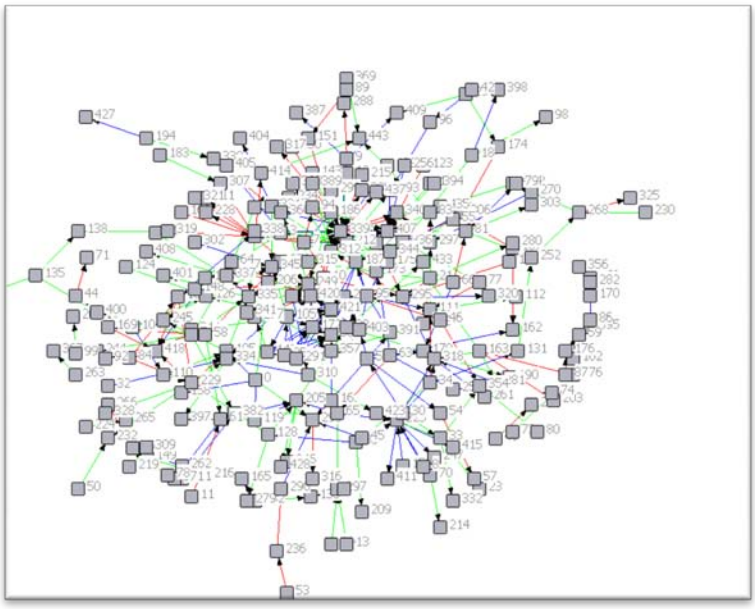


Figura 6.8: Red social generada con la totalidad de los representantes y tipos de relación.

Ahora podemos identificar aquellos sujetos y grupos de personas que pueden considerarse líderes o influyentes dentro de los grupos.

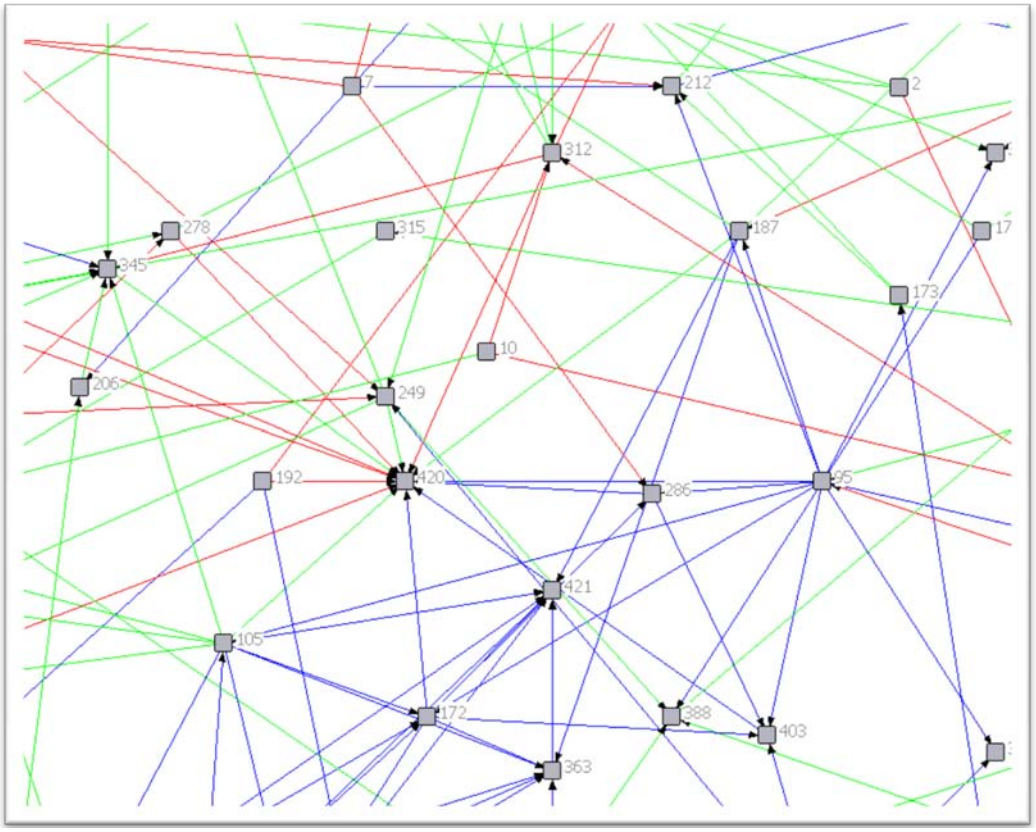


Figura 6.9: Detalle de Red social generada con la totalidad de los representantes y tipos de relación. En este caso podemos localizar el nodo con más incidencias detectándolo como influyente

O solo observar un tipo de relación en particular.

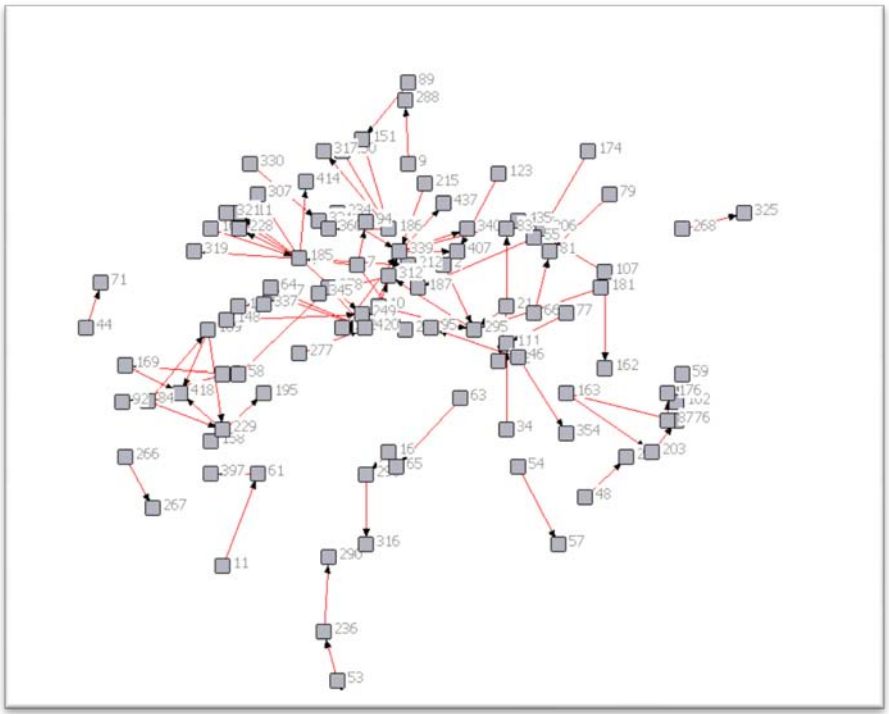


Figura 6.10: Red social generada con la totalidad de los representantes Solo con las relaciones de tipo políticas.

Dentro del análisis sobre los nodos podemos establecer criterios de selección para un atributo en particular de nodos.

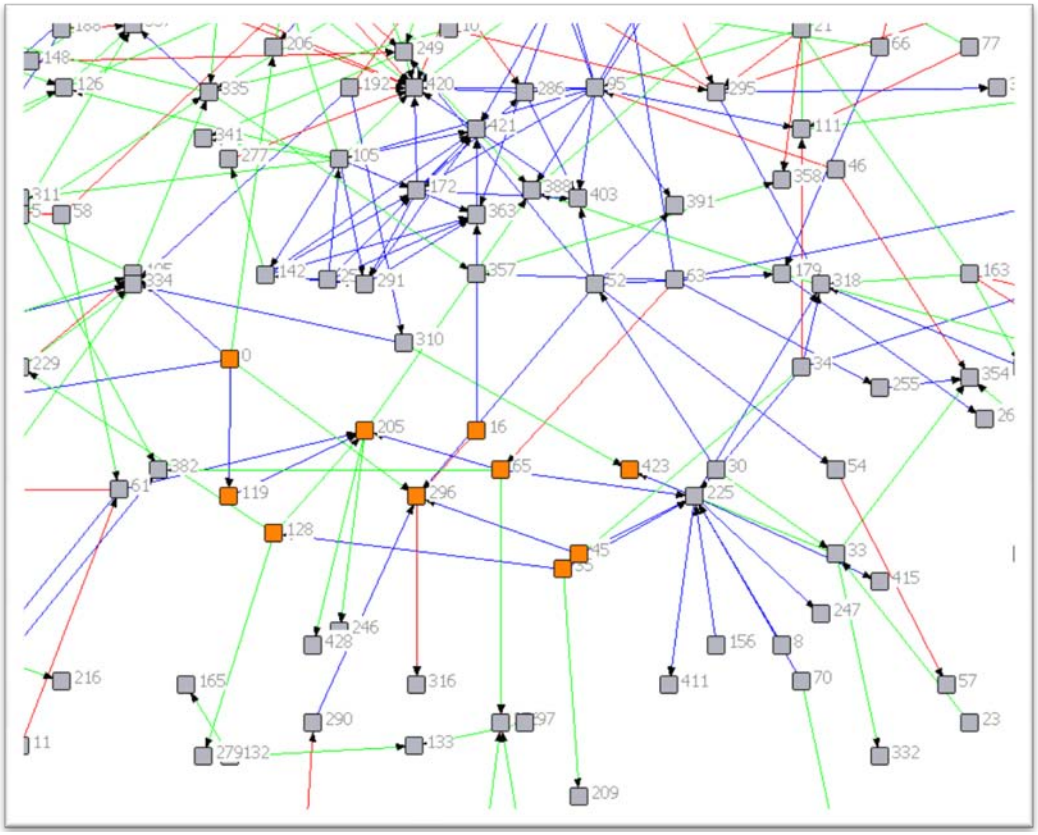


Figura 6.11: Detalle de Red social generada con la totalidad de los representantes destacando un atributo del nodo(representante) en particular, en este caso la postura política

El uso de nodos para la representación de redes sociales se convierte en una herramienta muy poderosa de análisis, cuyos reportes pueden ser de gran utilidad para el estudio y la reconstrucción de contextos sociales apoyando totalmente al estudio minucioso de biografías colectivas a partir de biografías individuales.

7. Resultados

7.1 Presentación de resultados

Durante todo el proceso del trabajo de minería de datos hemos superado diferentes etapas, que van desde la planeación y diseño de las bases de datos, la depuración de las mismas hasta la correcta selección y aplicación de algoritmos que nos permitan obtener información relevante para el historiador. A diferencia de aplicar minería de datos en otros sectores como el productivo o industrial, la presentación de los resultados de la misma debe ser por de más clara y sencilla de comprender, mostrando resultados y estadísticas concretas y útiles para el investigador histórico, por lo que el uso de gráficas, histogramas, tablas y diagramas debe ir acompañado de una breve explicación sin caer en tecnicismos.

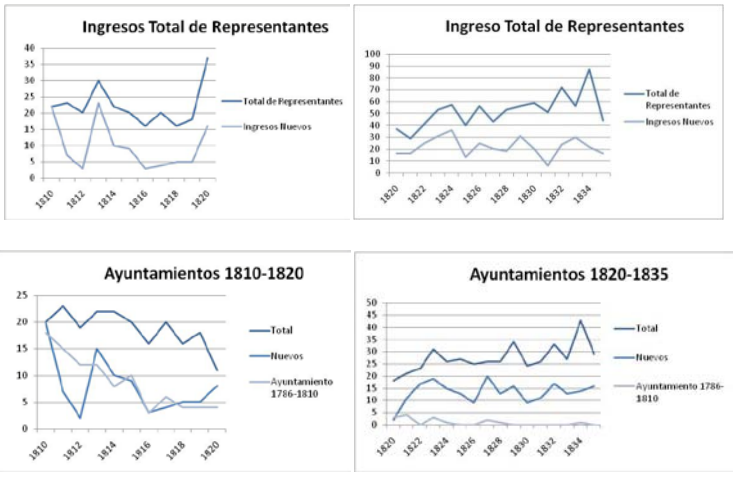
7.2. Resultados obtenidos y explicados.

7.2.1 Cargos

	Ayuntamiento Antiguo											Ayuntamiento															
	1810	1811	1812	1813	1814	1815	1816	1817	1818	1819	1820	1820	1821	1822	1823	1824	1825	1826	1827	1828	1829	1830	1831	1832	1833	1834	1835
Total	22	23	20	30	22	20	16	20	16	18	37	37	29	41	53	57	40	56	43	53	56	59	51	72	56	87	44
Ingreso Nuevos	22	7	3	23	10	9	3	4	5	5	16	16	16	25	31	36	13	25	20	18	31	20	6	24	30	22	16
	1810	1811	1812	1813	1814	1815	1816	1817	1818	1819	1820	1820	1821	1822	1823	1824	1825	1826	1827	1828	1829	1830	1831	1832	1833	1834	1835
Total Ayuntamiento	20	23	19	22	22	20	16	20	16	18	11	18	21	23	31	26	27	25	26	26	34	24	26	33	27	43	29
Ayuntamiento Nuevos	20	7	2	15	10	9	3	4	5	5	8	2	11	17	19	15	13	9	20	13	16	9	11	17	13	14	16
Ayuntamiento 1786-1810	18	15	12	12	8	10	3	6	4	4	4	3	4	0	3	1	0	0	2	1	0	0	0	0	0	1	0
	1810	1811	1812	1813	1814	1815	1816	1817	1818	1819	1820	1820	1821	1822	1823	1824	1825	1826	1827	1828	1829	1830	1831	1832	1833	1834	1835
Total Cortes	2	0	1	8	0	0	0	0	0	0	6	6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cortes Nuevos	2	0	1	8	0	0	0	0	0	0	6	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1810	1811	1812	1813	1814	1815	1816	1817	1818	1819	1820	1820	1821	1822	1823	1824	1825	1826	1827	1828	1829	1830	1831	1832	1833	1834	1835
Total Congreso	0	0	0	0	0	0	0	0	0	0	0	0	0	12	2	31	13	31	17	27	22	35	25	39	29	44	15
Congreso Nuevos	0	0	0	0	0	0	0	0	0	0	0	0	0	5	1	21	0	16	0	5	15	11	0	7	17	8	0
	1810	1811	1812	1813	1814	1815	1816	1817	1818	1819	1820	1820	1821	1822	1823	1824	1825	1826	1827	1828	1829	1830	1831	1832	1833	1834	1835
Total Provinciales	0	0	0	0	0	0	0	0	0	0	1	1	7	5	20	0	0	0	0	0	0	0	0	0	0	0	0
Provinciales Nuevos	0	0	0	0	0	0	0	0	0	0	0	0	5	3	11	0	0	0	0	0	0	0	0	0	0	0	0

Tabla 7.1: Tabla análisis y conteo de Cargos

Tabla 7.1 que mediante conteo muestra los ingresos de nuevos sujetos a Cargos públicos por año, también separados por la instancia de gobierno. Se muestran también los histogramas (Fig. 7.1) generados por la tabla anterior que nos permite deducir que entre el año 1812 y 1814 de primer periodo y 1834 del segundo hubo un aumento considerable en el registro de nuevos grupos políticos en cargos públicos.



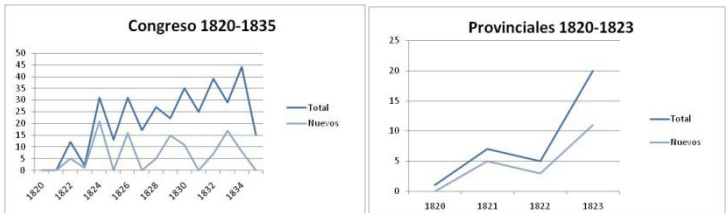


Figura 7.1: Histogramas de análisis de Cargos, que permiten al historiador detectar periodos clave

7.2.2 Oficios

Comerciante	131
Abogado	58
Hacendado	55
Capitán	54
Eclesiástico	53
Casateniente	31
Industrial	27
Teniente	25
Catedrático	23
Panadero	19
Abogado-de la audiencia Nacional	17
Abogado-de la real audiencia	17
Coronel	16
Teniente Coronel	15
Abogado de la audiencia del estado	14
Tocinero	14
Labrador	13
Molinero	13
Subteniente	13
Pulpero	11

Tabla 7.2: Actividades y oficios de RepresentantesP

Los Representantes populares además de su cargo público tenían otras actividades u oficios, los cuales pueden permitir al historiador conocer su estatus económico y de acuerdo a la naturaleza del oficio o actividad desempeñada formularse hipótesis respecto a las relaciones posibles con otros personajes. La tabla 7.2 nos muestra que la mayoría de los representantes se dedicaba al comercio o al ejercicio de las leyes. El siguiente gráfico representa la distribución por categorías de los oficios.

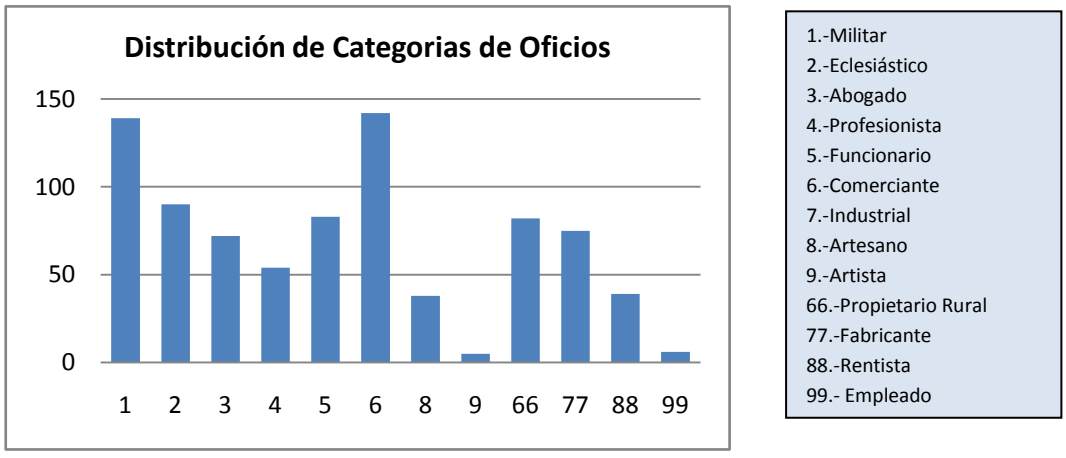


Figura 7.2: Distribución de categorías por oficios.los oficios están categorizados de acuerdo a la tabla

Tabla 7.2: Actividades y oficios de RepresentantesP

Algunos representantes no solo tenían una actividad aparte de su cargo público, sino que combinaban otras actividades, se obtuvieron las combinaciones más populares dentro del contexto.

0166----	18	0677----	4
0601----	8	77------	4
04------	6	0108----	3
08------	6	0202----	3
7706----	6	020202--	3
0305----	5	0266----	3
05------	5	0404----	3
0503----	5	0505----	3
0666----	5	050505--	3
0104----	4	0688----	3
0204----	4	0806----	3
020402--	4	77016677	3
0606----	4	770601--	3

Tabla 7.3: Combinaciones más populares de oficios en categorías

De esta forma podemos concluir que muchos Representantes además de tener un grado militar poseían un comercio o una propiedad rural.

7.2.3 Posturas

Mediante gráficas históricas se representan los cambios de postura de los representantes políticos a lo largo del periodo

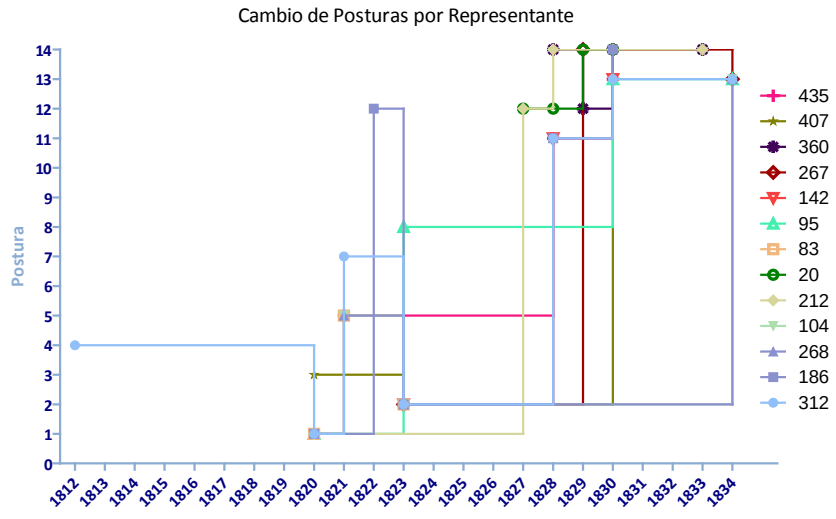


Figura 7.3: Gráfico histórico del cambio de posturas de los representantes

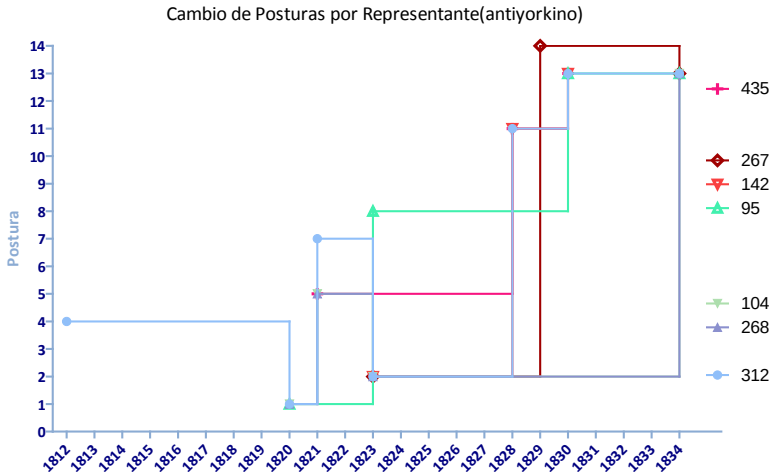


Figura 7.4: Gráfico histórico del cambio de posturas de los representantes antiyorkinos

Únicamente se muestran los de la corriente antiyorkina

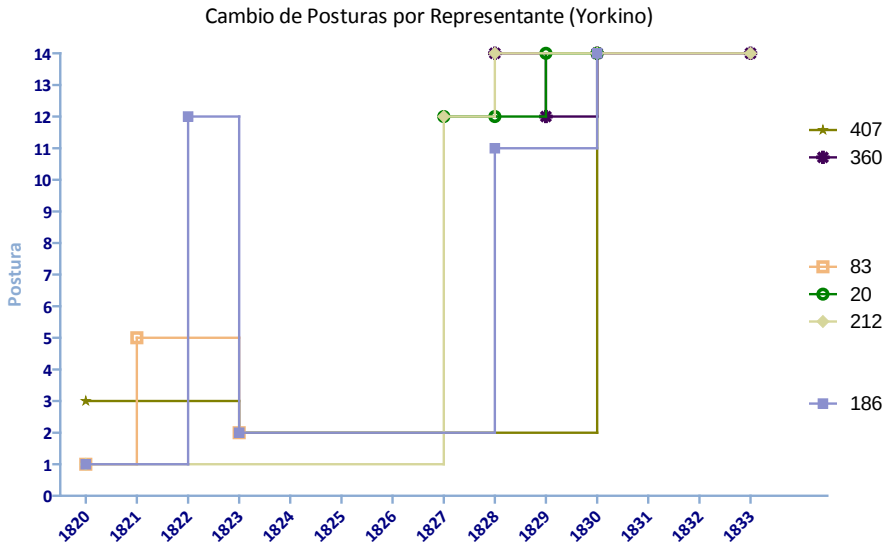


Figura 7.5: Gráfico histórico del cambio de posturas de los representantes yorkinos

En este caso solo los yorkinos.

7.2.4 Representantes

Se pudo deducir una relación existente entre el lugar de origen y la edad de los representantes populares, con el número de cargos ocupados así mismo de la corriente política y su capital.

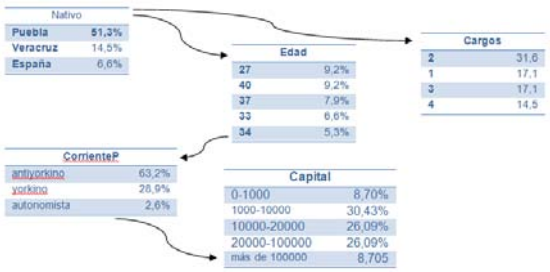


Figura 7.5: Relaciones entre atributos lugar de origen y edad de representantes así como el número de cargos ocupados

7.2.5 Relaciones

Figura 7.6: Representación gráfica de todas las relaciones del universo de estudio. Permite al historiador detectar los diferentes grupos formados a partir de los 3 tipos de relación y de cierta forma analizar su comportamiento para establecer movimientos políticos

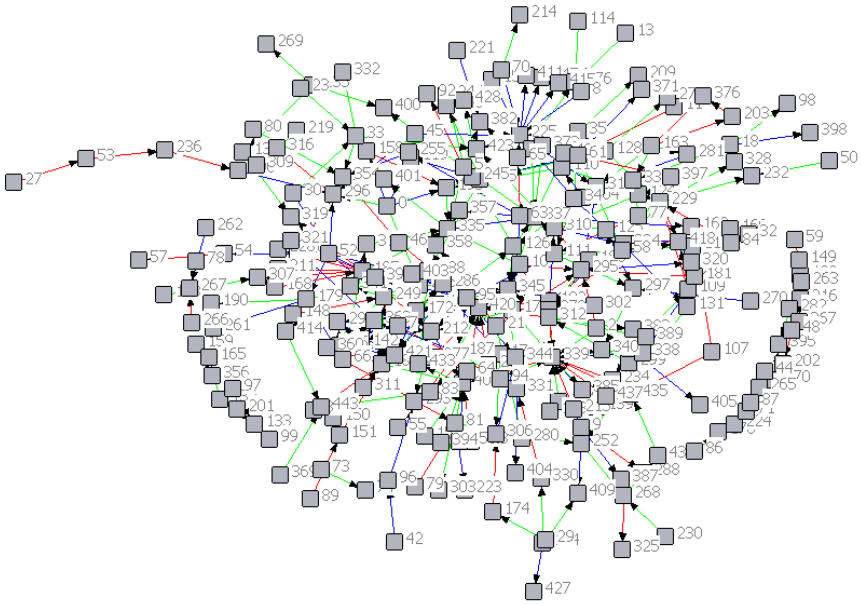
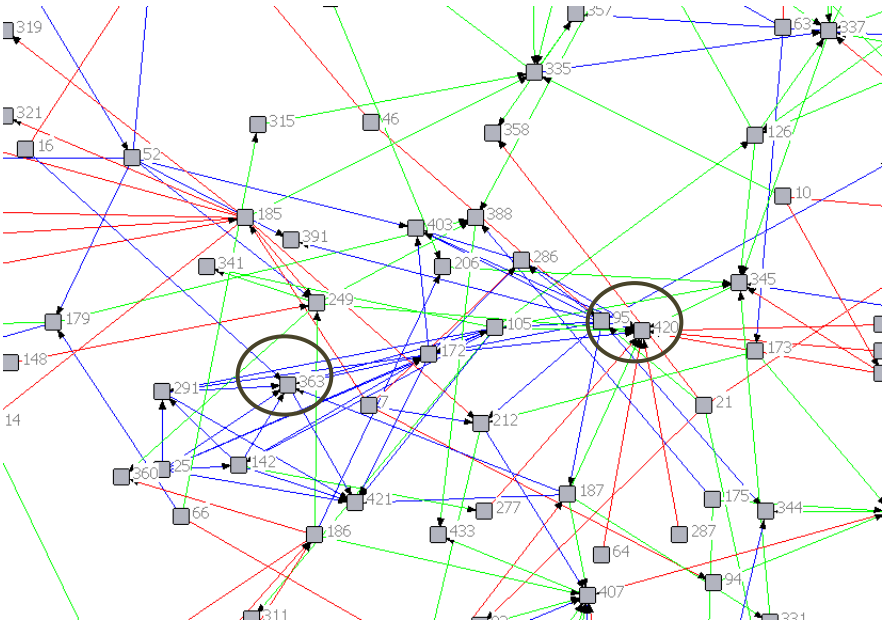


Figura 7.7: Detección de los Representantes más influyentes y del grupo central de relaciones. La gráfica de la red social nos permite conocer aquellos nodos (representantes) más influyentes mediante la aparición diversas conexiones tanto políticas como económicas como políticas.



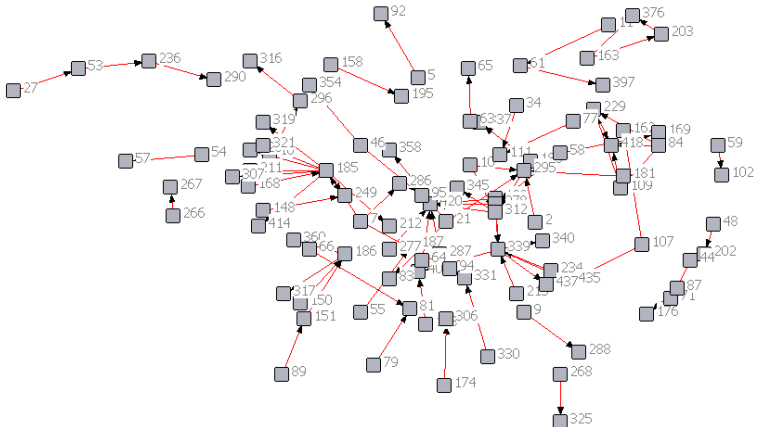


Figura 7.8: Gráfico de relaciones de tipo político. Muestra los grupos formados únicamente a través de las relaciones políticas detectadas en la investigación y asi detectar grupos en particular.

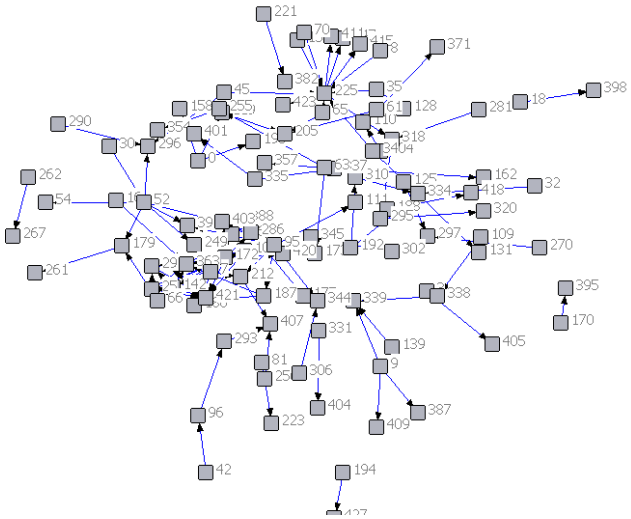


Figura 7.9: Gráfico de relaciones de tipo económico. Las relaciones económicas, como préstamos, fiadores, relaciones comerciales y demás, de alguna manera pueden generar una relación política y una convergencia de grupos, por lo que este gráfico permite al historiador conocerlos.

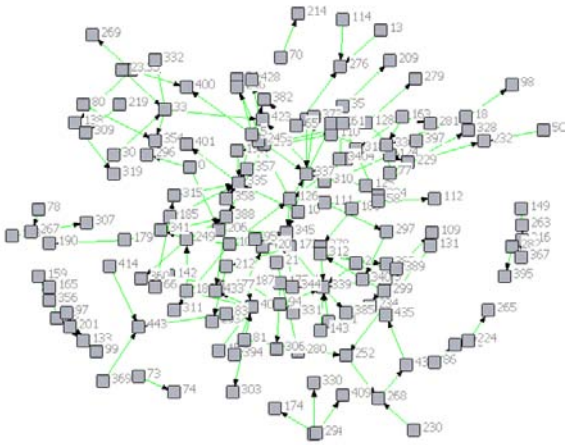


Figura 7.10: Gráfico de relaciones de tipo familiar. Las relaciones de tipo familiar permiten tratar de detectar la formación de grupos políticos atreves de nexos familiares

8. Conclusiones

8.1 Conclusiones

La prosopografía como método de investigación, particularmente en el caso de la historia política, es una herramienta eficaz para la reconstrucción de contextos históricos, tiene una relación que puede hacerse muy estrecha desde el enfoque de obtener conocimiento a partir de seleccionar y analizar exhaustivamente y de manera, ordenada, clara y concreta los registros de en el caso de la primera individuos, en el caso de la segunda, una gama muy amplia de datos.

La investigación histórica es una rama de las ciencias sociales que puede ser auxiliada perfectamente por la informática siempre y cuando exista la apertura de los especialistas de ambas disciplinas de generar los puentes de colaboración y estudio interdisciplinario. Implementar sistemas de información, crear bases de datos, analizarlas y aprovechar las diferentes soluciones ya existentes para el apoyo de la investigación en ciencias sociales significará ahorrar, tiempo y esfuerzo además de aprovechar el trabajo ya realizado por otros investigadores.

La aplicación de la minería de datos ha sido significativa en la toma de decisiones de empresas o industrias, aquellos que son capaces de obtener conocimiento y perspectivas claras a partir del manejo y análisis de la información que generan están ahorrando grandes cantidades de tiempo y aumentando su panorama, este mismo modelo y aplicación puede darse también en la investigación social.

La minería de datos es todo un conjunto de herramientas que concentra diferentes disciplinas computacionales, como la estadística, la recuperación de datos, la computación paralela, el aprendizaje automático o la gestión de bases de datos, como tal, es un universo muy amplio que obliga al especialista en recuperación del conocimiento a mantener una disciplina de estudio y una apertura total al aprendizaje constante de diferentes algoritmos.

Desarrollar un proyecto de minería de datos siguiendo perfectamente cada uno de los pasos que implica, genera la seguridad suficiente para confiar en los resultados obtenidos. La aplicación de los algoritmos de minería de datos aunada al uso de aplicaciones estadísticas genera resultados comprensibles y ricos de conocimiento para el investigador histórico, quien debe poder leer y entender los resultados generados por el sistema utilizando un lenguaje claro e intuitivo para el despliegue de resultados, auxiliándose en diagramas, consultas o histogramas.

8.2 Trabajo Futuro

Existe una gran cantidad de disciplinas dentro de las ciencias sociales en las que la aplicación de técnicas de recuperación y análisis de información puede ser muy útil, actualmente se desarrollan, aunque lentamente, herramientas computacionales para el trabajo con ellas.

La historia política, la historia social o la historia económica son disciplinas en las que el especialista puede trabajar arduamente auxiliando y generando modelos de implementación de técnicas de minería de datos para la reconstrucción y clarificación de datos del pasado.

Las redes sociales vistas desde el punto de vista sociológico son por ejemplo una magnífica oportunidad donde las ciencias computacionales y exactas tienen un gran campo de aplicación.

El especialista en recuperación de información, debe acercarse al investigador social y conocer sus necesidades, inmiscuirse en el tipo de información que este genera y en el posible conocimiento que se puede obtener a partir de ellos.

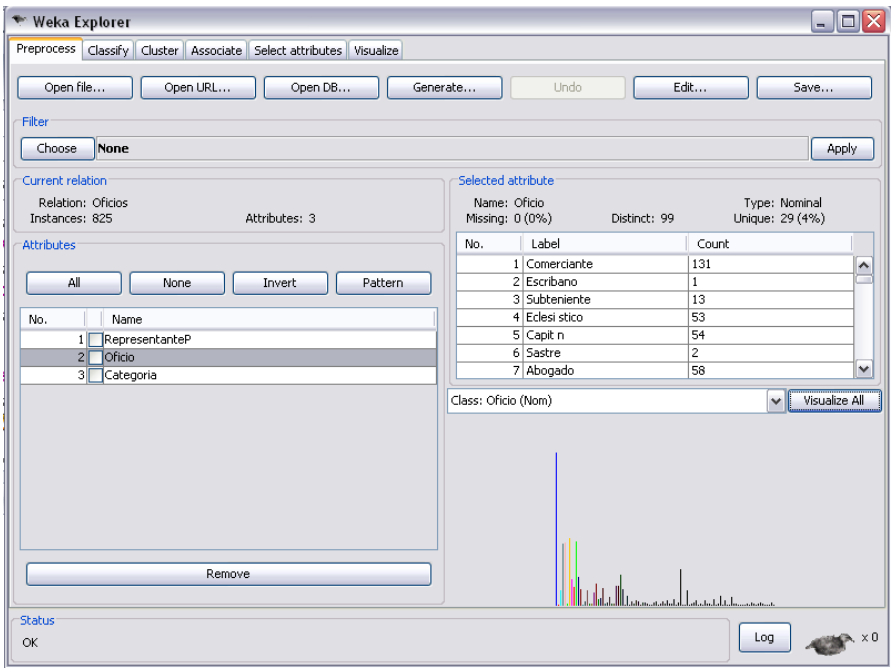
Los sistemas de información histórica son una herramienta que no ha terminado de despegar, por lo que la correcta aplicación y desarrollo de los mismos, puede significar el despegue total de la aplicación informática en las ciencias sociales.

Es necesario también desarrollar y comprender nuevas técnicas de minería de datos particularmente en el manejo de información cualitativa, por lo que el aprendizaje automático y las herramientas deben ser cada día más poderosas.

Apéndice A. Herramientas para minería de datos y sih

Weka.

El paquete Weka[4] contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades. La versión original de Weka fue un front-end en TCL/TK para modelar algoritmos implementados en otros lenguajes de programación, más unas utilidades para preprocesamiento de datos desarrolladas en C para hacer experimentos de aprendizaje automático. Esta versión original se diseñó inicialmente como herramienta para analizar datos procedentes del dominio de la agricultura,[5] [6] pero la versión más reciente basada en Java (WEKA 3), que empezó a desarrollarse en 1997, se utiliza en muchas y muy diferentes áreas, en particular con finalidades docentes y de investigación.



Más información en:
http://es.wikipedia.org/wiki/Weka_(aprendizaje_autom%C3%A1tico)

SAS Enterprise Minner

Enterprise Minner simplifica el proceso de minería de datos de alta precisión para crear modelos descriptivos y predictivos basados en el análisis de grandes cantidades de datos de toda la empresa. Agiliza el proceso de minería de datos desde el acceso al modelo de datos y al modelo de despliegue mediante la aplicación de tareas necesarias dentro de una sola solución integrada, ofreciendo flexibilidad para una eficiente colaboración de grupos de trabajo.

Múltiples interfaces de

* Procesamiento escalable

* Preparación de datos, resúmenes y exploración
* modelos avanzados predictivos y descriptivos
* Modelo de negocios basado comparaciones, presentación de informes y la gestión
* Porceso avanzado de automatizaciónautomatizado

Más información en: http://www.sas.com/technologies/analytics/datamining/miner/

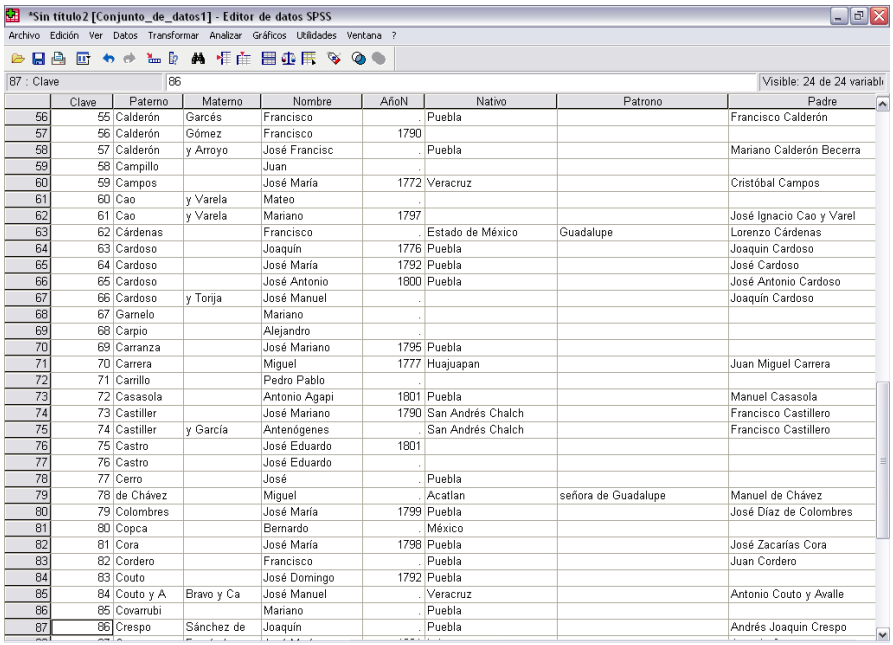
SPSS

Statistical Package for the Social Sciences (SPSS) es un programa estadístico informático muy usado en las ciencias sociales y las empresas de investigación de mercado.

El sistema de módulos de SPSS, como los de otros programas (similar al de algunos lenguajes de programación) provee toda una serie de capacidades adicionales a las existentes en el sistema base. Algunos de los módulos disponibles son:

- Modelos de Regresión
- Modelos Avanzados
    - Reducción de datos: Permite crear variables sintéticas a partir de variables colineales por medio del Análisis Factorial.
    - Clasificación: Permite realizar agrupaciones de observaciones o de variables (*cluster analysis*) mediante tres algoritmos distintos.
    - Pruebas no paramétricas: Permite realizar distintas pruebas estadísticas especializadas en distribuciones no normales.
- Tablas: Permite al usuario dar un formato especial a las salidas de los datos para su uso posterior. Existe una cierta tendencia dentro de los usuarios y de los desarrolladores del software por dejar de lado el sistema original de TABLES para hacer uso más extensivo de las llamadas CUSTOM TABLES.
- Tendencias

- Categorías: Permite realizar análisis multivariados de variables normalmente categorías. También se pueden usar variables métricas siempre que se realice el proceso de recodificación adecuado de las mismas.
- Análisis Conjunto: Permite realizar el análisis de datos recogidos para este tipo especifico de pruebas estadísticas.
- Mapas: Permite la representación geográfica de la información contenida en un fichero (descontinuado para SPSS 16).
- Pruebas Exactas: permite realizar pruebas estadísticas en muestras pequeñas.
- Análisis de Valores Perdidos: Regresión simple basada en imputaciones sobre los valores ausentes.
- Muestras Complejas: permite trabajar para la creación de muestras estratificadas, por conglomerados u otros tipos de muestras.
- SamplePower (cálculo de tamaños muestrales)
- Árboles de Clasificación: Permite formular árboles de clasificación y/o decisión con lo cual se puede identificar la conformación de grupos y predecir la conducta de sus miembros.
- Validación de Datos: Permite al usuario realizar revisiones lógicas de la información contenida en un fichero.sav. y obtener reportes de los valores considerados extraños. Es similar al uso de sintaxis o scripts para realizar revisiones de los ficheros. De la misma forma que estos mecanismos es posterior a la digitalización de los datos.



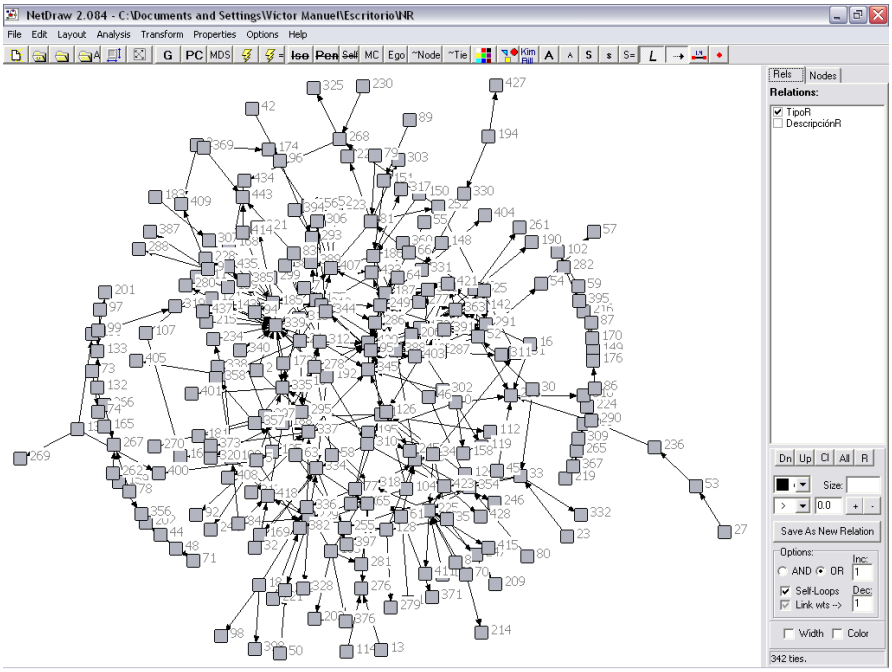
	Clave	Paterno	Materno	Nombre	AñoN	Nativo	Patrono	Padre
56	55	Calderón	Garcés	Francisco	.	Puebla		Francisco Calderón
57	56	Calderón	Gómez	Francisco	1790			
58	57	Calderón	y Arroyo	José Francisc	.	Puebla		Mariano Calderón Becerra
59	58	Campillo		Juan				
60	59	Campos		José María	1772	Veracruz		Cristóbal Campos
61	60	Cao	y Varela	Mateo	.			
62	61	Cao	y Varela	Mariano	1797			José Ignacio Cao y Varel
63	62	Cárdenas		Francisco	.	Estado de México	Guadalupe	Lorenzo Cárdenas
64	63	Cardoso		Joaquín	1776	Puebla		Joaquin Cardoso
65	64	Cardoso		José María	1792	Puebla		José Cardoso
66	65	Cardoso		José Antonio	1800	Puebla		José Antonio Cardoso
67	66	Cardoso	y Torija	José Manuel	.			Joaquín Cardoso
68	67	Garnelo		Mariano	.			
69	68	Carpio		Alejandro	.			
70	69	Carranza		José Mariano	1795	Puebla		
71	70	Carrera		Miguel	1777	Huajuapan		Juan Miguel Carrera
72	71	Carrillo		Pedro Pablo	.			
73	72	Casasola		Antonio Agapi	1801	Puebla		Manuel Casasola
74	73	Castiller		José Mariano	1790	San Andrés Chalch		Francisco Castillero
75	74	Castiller	y García	Antenógenes	.	San Andrés Chalch		Francisco Castillero
76	75	Castro		José Eduardo	1801			
77	76	Castro		José Eduardo	.			
78	77	Cerro		José	.	Puebla		
79	78	de Chávez		Miguel	.	Acatlan	señora de Guadalupe	Manuel de Chávez
80	79	Colombres		José María	1799	Puebla		José Díaz de Colombres
81	80	Copca		Bernardo	.	México		
82	81	Cora		José María	1798	Puebla		José Zacarías Cora
83	82	Cordero		Francisco	.	Puebla		Juan Cordero
84	83	Couto		José Domingo	1792	Puebla		
85	84	Couto y A	Bravo y Ca	José Manuel	.	Veracruz		Antonio Couto y Avalle
86	85	Covarrubi		Mariano	.	Puebla		
87	86	Crespo	Sánchez de	Joaquín	.	Puebla		Andrés Joaquin Crespo

Más información en: http://es.wikipedia.org/wiki/SPSS

NetDraw

Esta es una aplicación gratuita diseñada para el análisis y diseño de redes sociales, entre las herramientas con que cuenta el software se pueden contar las siguientes:

- Puede manejar relaciones multiples entre los nodos de la red.
- Le permite asignar valores de importancia a los nodos de la red, así como atributos, los cuales le permiten formar subgrupos y hacer una mejor representación del modelo.
- Incluye un set de procedimientos de nálisis comúnmente usados en este tipo de estudios, tales como identificación de nodos ailados, componentes, k-cores, entre otros.
- Es posible leer datos de representación a través de la herramienta, la cual es capas de entender el lenguaje de Davis, así como Gardner and Gardner data y crear automáticamente representaciones a través de este tipo de datos.
- La aplicación cuenta con una interfaz gráfica más o menos sencilla de utilizar, la cuál además es configurable y permite exportar los datos creados a distintos formatos para su posterior uso.

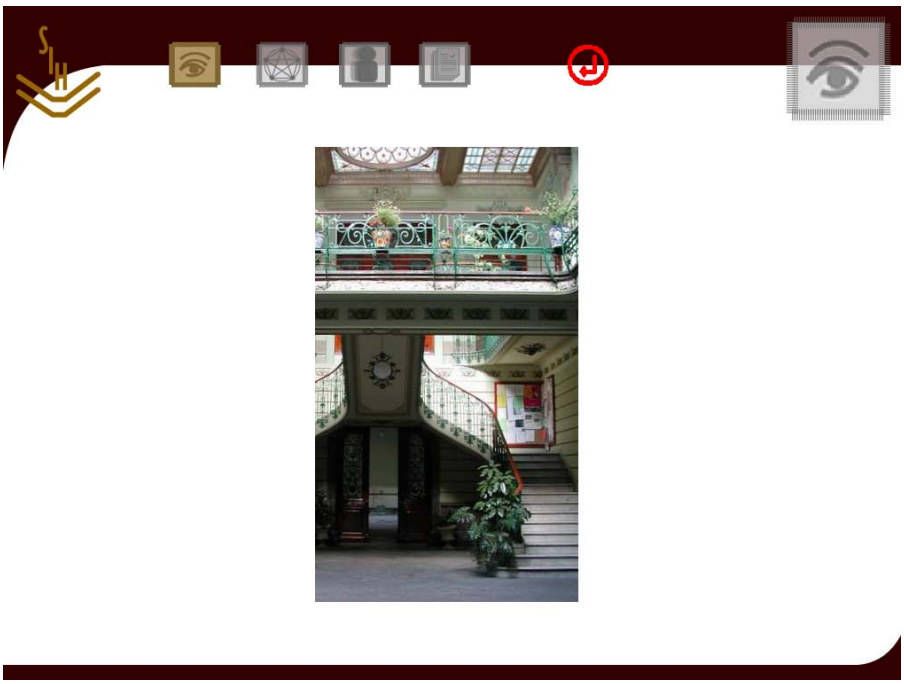


Más información en: http://www.analytictech.com/Netdraw/netdraw.htm

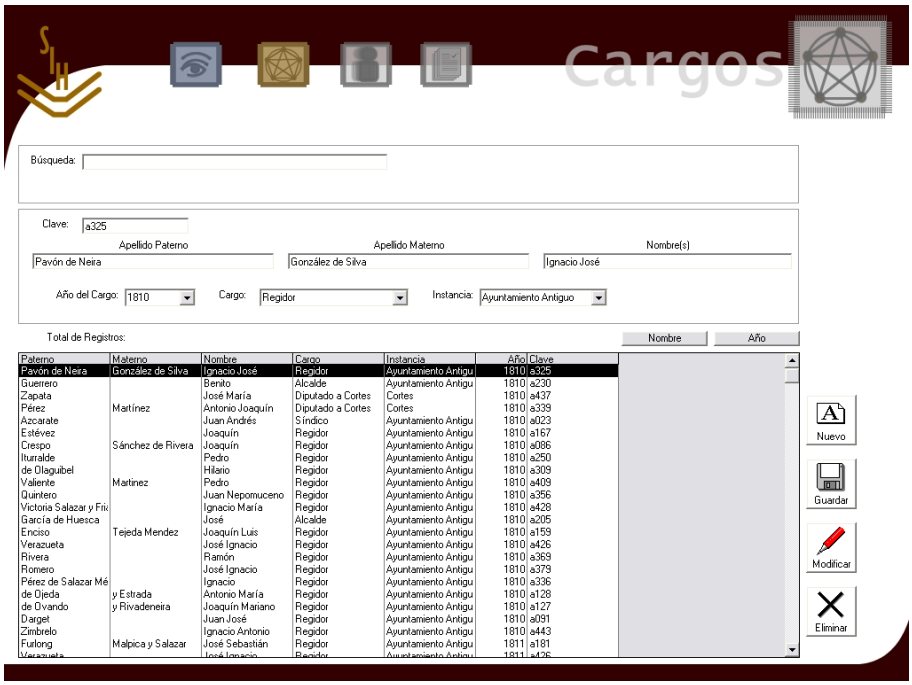
Apendice B SIH creado para el proyecto.

Durante la primera parte de la investigación del presente trabajo y dentro de la estancia con el investigador para conocer en práctica los métodos de investigación histórica, se generó un sistema de captura, consulta y análisis de los datos, para garantizar que los datos fueran introducidos de la mejor manera posible, además de ofrecer soluciones de consulta y análisis básicas para el historiador, sin embargo muchos de los resultados expuestos en este trabajo fueron producto de un posterior análisis y refinamiento.

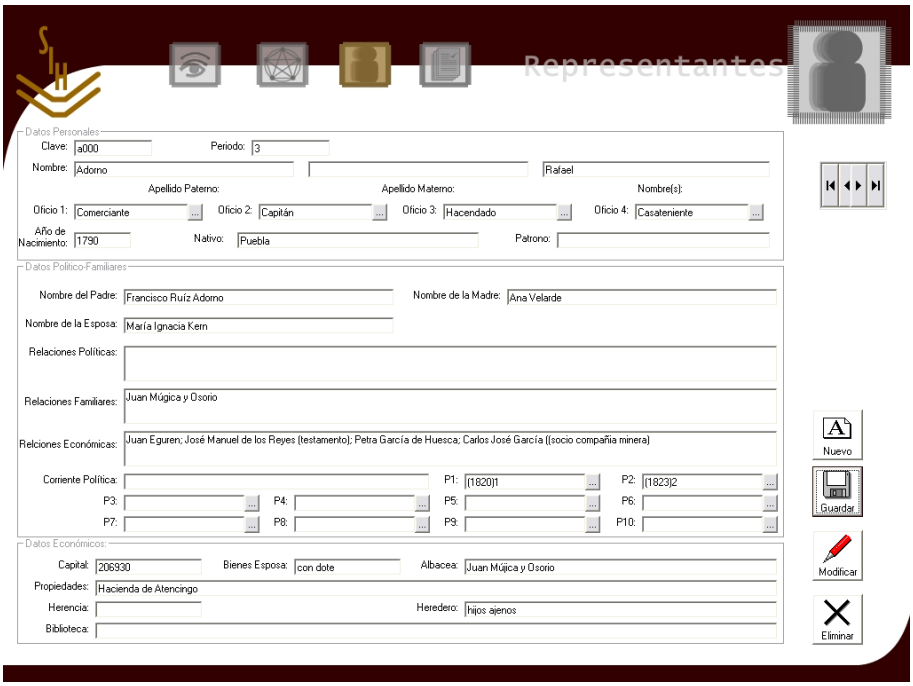
A continuación se describe la interfaz implementada para el uso del historiador y sus auxiliares.



Pantalla Principal con acceso a los 3 principales accesos. La captura de cargos, la captura de representantes y el área de reportes y filtros.

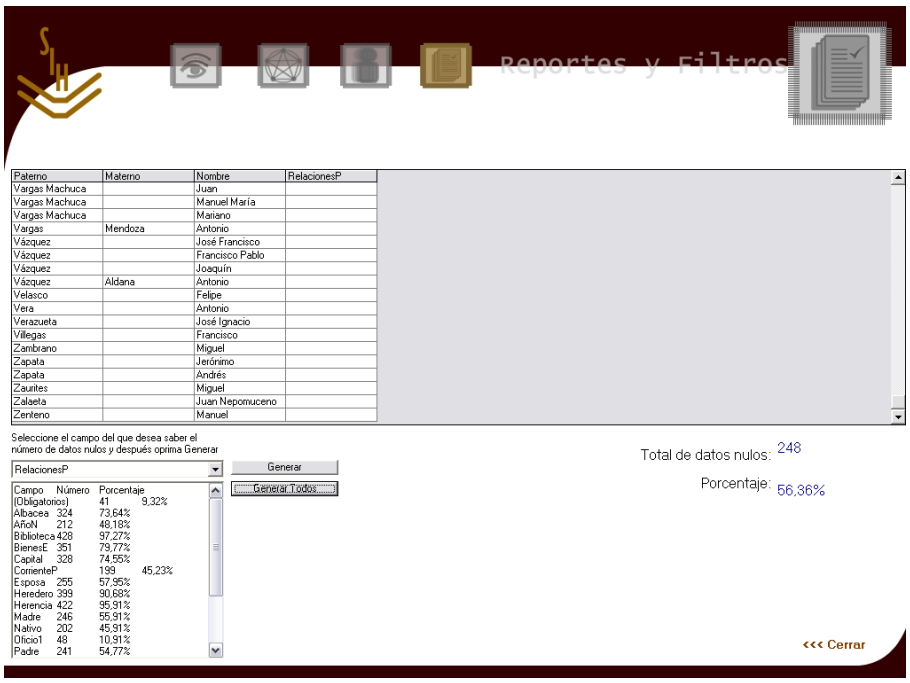


Área de captura y modificación de la tabla de cargos.

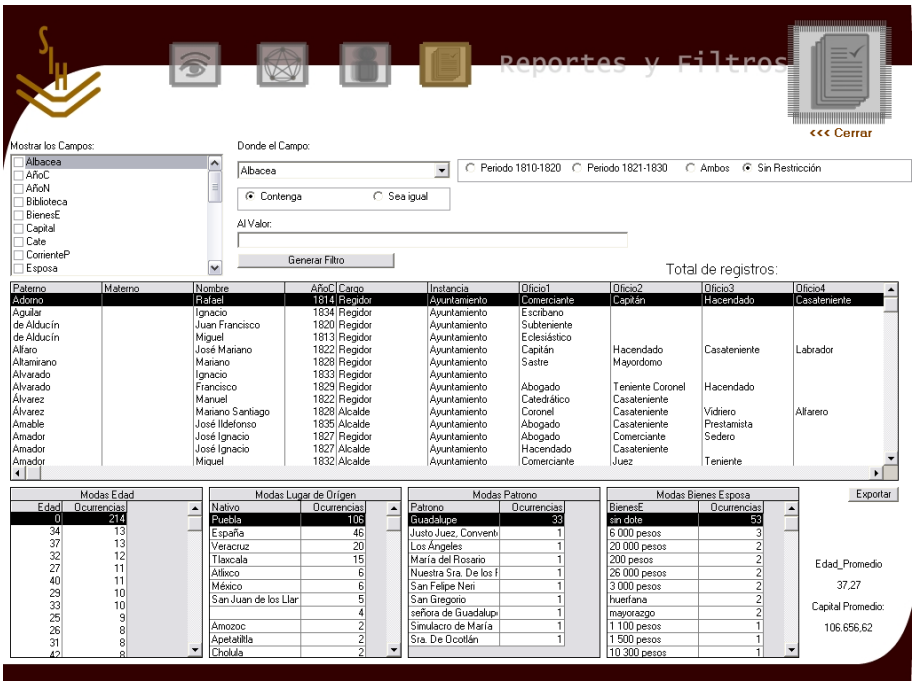


Área de captura y modificación de la tabla de Representantes.

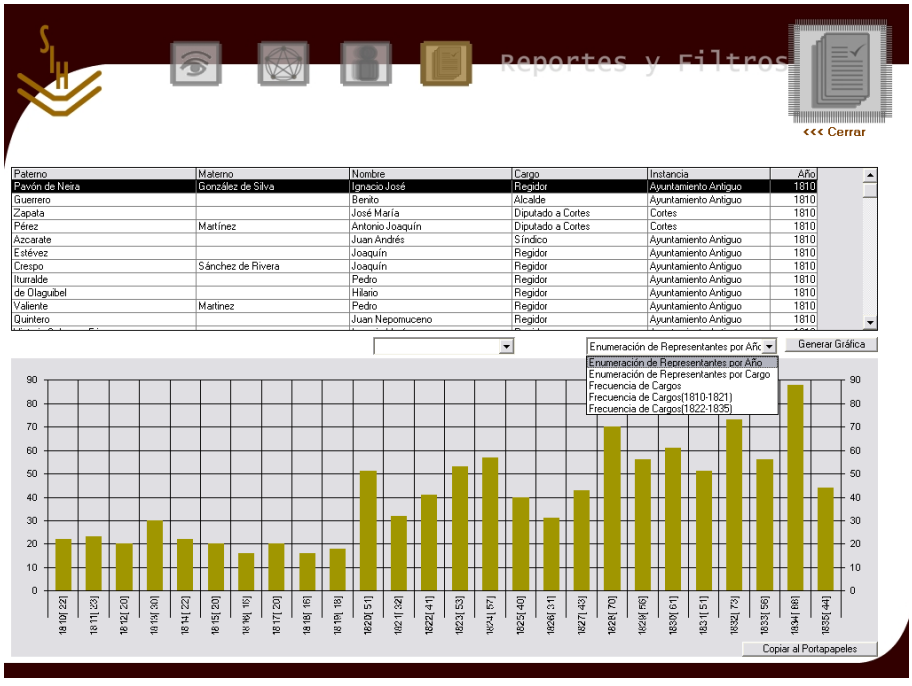
Sección de Reportes y filtros, ofrece un menú con las diferentes opciones de consulta.



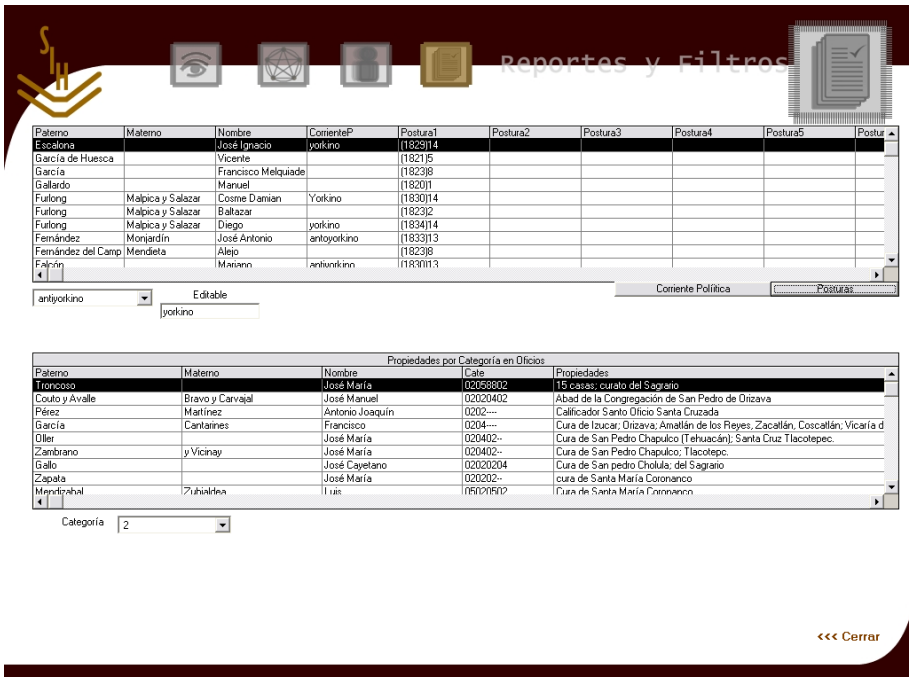
Consulta de los datos nulos de toda la base de datos, que permite al investigador tener presente el índice de efectividad de los datos de los Representantes populares.



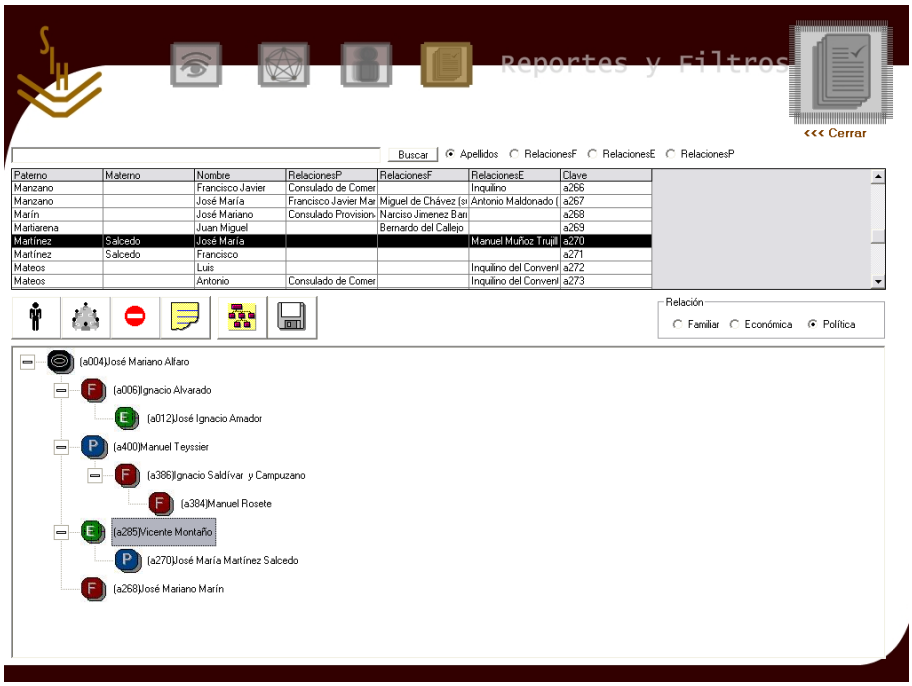
Esta sección ofrece diversas posibilidades de consulta, selección y ordenamiento de los datos de los Representantes, permite definir los campos a mostrar, el periodo y un filtro sobre algún atributo en particular, además muestras las medias y modas de los datos cuantitativos.



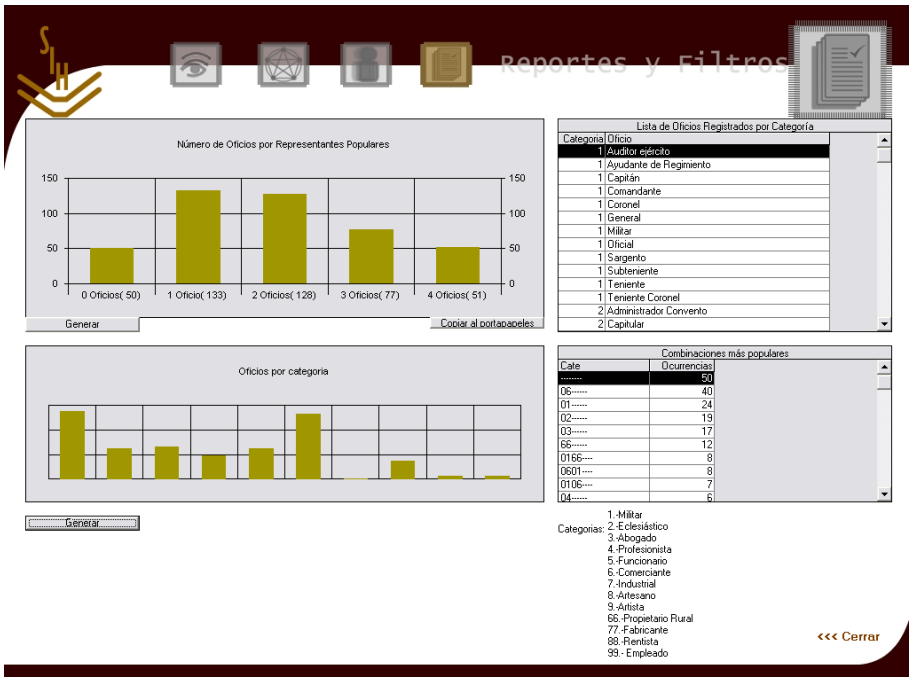
Esta ventana permite visualizar gráficamente diversas estadísticas básicas respecto a los cargos de los representantes.



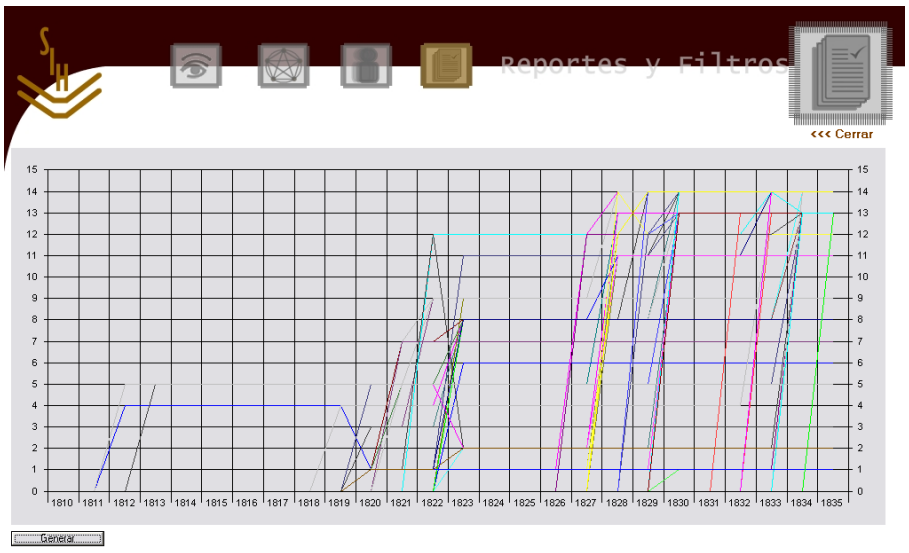
Consulta de posturas y corrientes políticas.



Esta herramienta permite al historiador hacer búsquedas de Representantes y construir manualmente árboles de relación entre ellos, permite guardarlos como archivo y abrirlos posteriormente.



Esta ventana permite visualizar gráficamente diversas estadísticas básicas respecto a los oficios de los representantes



Gráfica que se autogenera describiendo el cambio de posturas de todos los representantes involucrados, al profundizar sobre el tema, se opto por particionar y optimizar los grupos de personas utilizando graficadores.

1) **Alía** F.; *El Historiador y Las Bases De Datos;* en La historia en una nueva frontera; Ediciones de la Universidad de Castilla-La Mancha, 2000.
2) **Barros**, C.; *Historia a Debate;* Actas del II Congreso Internacional, tomo II: Nuevos paradigmas, Historia a Debate, Santiago de Compostela, 2000
3) **Cabena** P., **Hadjinian** P., **Standler** R., **Verhees** J., **Zanasi** A. ; *Discovering Data Mining. From Concept to Implementation*; Prentice Hall, 1998.
4) **Carr** E.H.; *Qué es la historia*; Seix Barral, 1970.
5) **Dedieu**, J.P.; Un instrumento para la historia social: la base de datos Ozanam., en Cuadernos de Historia Moderna, Universidad Complutense de Madrid, nº 24, año 2000
6) **Fayyad** U.M., **Piatetsky-Shapiro** G., **Smyth** P., **Uthurusamy** R.; *Advances in Knowledge Discovery and Data Mining;* AAAI Press, 1996.
7) **Fernández** F.; *La Historia Moderna y Nuevas Tecnologías De La Información y Las Comunicaciones*; Cuadernos de Historia Moderna, Nº 24, pág. 11-31, 2000.
8) **Galiano** F.; *Apuntes De Minería De Datos y Descubrimiento Del Conocimiento*; Universidad de Granada, España, 2000.
9) **García** F.J.; *Información Digital e Investigación Histórica: Una Aproximación*; Biblios Nº 9 , 2001.
10) **García** F.J.; *Sistemas De Información Histórica (SIH): La Documentación Al Servicio Del Pasado*; Anuario de Biblioteconomía, Documentación e Información Nº. 2001-2002, pags. 75-93 , 2003.
11) **Genet** J.P.; *Informatique Et Histoire*; Le Bulletin De L'EPI N° 49, pág. 133-144, 1988.
12) **Guerra** F. X.; *México Del Antiguo Régimen a La Revolución*; Fondo de Cultura Económica; México, 1993.
13) **Han** J., **Kamber** M.; *Data Mining: Concepts and Techniques*; Morgan Kaufmann; 2006.
14) **Hernández** J., **Ramírez** M.J., **Ferri** C.; *Introducción a La Minería De Datos*; Editorial Pearson, 2004.
15) **Larose** D.T.; *Data Mining Methods and Models*; Wiley-IEEE Press, 2006.
16) **López** J.; *Las Bases De Datos Históricas*; Biblios Nº 9 , 2001.
17) **Michalski** R.S., **Bratko** I., **Kubat** M.; *Machine Learning and Data Mining*; Johb Wiley and Sons, 1998.
18) **Molina** R.; *De La Utilidad y Los Inconvenientes De La Informática Para La Historia*; Tiempos modernos: Revista Electrónica de Historia Moderna, Nº. 7, 2002-2003.
19) **Olagüe de Ros** G.; *De Las Vidas Ejemplares a Las Biografías Colectivas De Los Médicos. Una Perspectiva Crítica*; Asclepio-Vol. LVII-1, Pág. 135-148, 2005.
20) **Pérez** C., **Santin** D.; *Data Mining. Soluciones Con Enterprise Miner*; Editorial Rama, 2006.

21) **Rousseau** I.; *La Prosopografía: ¿Un Método Idóneo Para El Estudio Del Estado?*; Revista Mexicana de Sociología, Vol 52, Nº 3, pág. 237-247, 1990.

22) **Sánchez** E.; *Fuentes Para Una Prosopografía De Los Mercaderes Novohispanos: El Caso De Cuernavaca y Cuautla De Amilpas (Morelos) En El Siglo XVIII;* América Latina en la historia económica, Nº 18, 2002.

23) **Sánchez** M.I.; *Análisis De Redes Sociales: Una Herramienta En Manos De Los Historiadores En La Historia En Una Nueva Frontera;* Ediciones de la Universidad de Castilla-La Mancha, 2000.

24) **Silva** J.; *Reseña De "Grandeur Et Misére De L'office. Les Officers De Finances De Nouvelle-Espagane XVIIE-XVIIIE Siécles De. Sobre Michel Bertrand*; Historia Mexicana, octubre-diciembre, año/vol. LII, Nº 002; 2002.

25) **Sordo** R.; *La Historia Política Del Siglo XIX: De La "Historia Tradicional" A La "Nueva Historia"*; en Cincuenta años de Investigación Histórica en México; Universidad Nacional Autónoma de México, Instituto de Investigaciones Históricas/Universidad de Guanajuato, Serie Historia Moderna y Contemporánea 29, 1998.

26) **Stone** L.; *El pasado y el presente*; Fondo de Cultura Económica; México, 1986.

27) **Téllez** D.; *D. Ricardo Wall: De La Biografía, La Narratividad, La Prosopografía, El Hipertexto y Otras Especies*; Tiempos modernos: Revista Electrónica de Historia Moderna, Vol. 3, Nº 7; 2002.

28) **Trejo** Z.; *Del Fusil a La Palabra: Redes Sociales, Facciones y Liberalismo En Los Congresos Sonorenses, 1847-1876*; Boletín Portales del Colegio de Sonora, Año 5, Nº 195, 2006.

29) **Van Dalen** D. B., **Meyer** W.J.; *Estrategia De La Investigación Histórica*; Manual de técnica de la investigación educacional; 1978.

30) **Velarde** M.A. Minería de Datos. Una Introducción; Instituto Tecnológico de Aguascalientes; 2003,

31) **Witten** I.H., **Frank** E.; *Data Mining : Practical Machine Learning Tools And Techniques*; 2ª edición; Morgan Kaufmann; 2005