UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

APORTACIONES AL ANÁLISIS Y VISUALIZACIÓN INFORMÉTRICA CON LA RED NEURONAL SOM

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

P R E S E N T A:

ELIO ATENÓGENES VILLASEÑOR GARCÍA

Director de Tesis:

DR. HUMBERTO ANDRÉS CARRILLO CALVET

Facultad de Ciencias, UNAM

Ciudad Universitaria, CD. MX. Mayo, 2018.

A Martha y Edahi.

Prefacio

La tecnología moderna nos ha permitido disponer de enormes capacidades computacionales y de repositorios capaces de almacenar grandes volúmenes de información. Esto ha tenido como consecuencia que la minería de datos y el descubrimiento de conocimiento en bases de datos (KDD) constituyan temas estratégicos de actualidad.

Paradigmas emergentes de la inteligencia computacional, como las redes neuronales, han contribuido con algoritmos y métodos efectivos para el desarrollo de las complejas tareas analíticas que requiere el descubrimiento y visualización del conocimiento intrínseco en los datos. Entre las redes neuronales más usadas en la minería de datos se encuentran las redes de Kohonen basadas en el algoritimo SOM - por sus siglas en inglés: *Self-Organizing Maps*. Estas redes son muy apreciadas porque permiten procesar eficientemente grandes volúmenes de datos, de alta dimensionalidad y alta complejidad estructural (datos con componentes numéricas y categóricas), de los cuales no es necesario tener conocimiento a priori ya que el proceso de aprendizaje de la red neuronal es no-supervisado.

Solamente en la base de Scopus se tienen registrados 13,870 publicaciones científicas cuyos registros bibliográficos son recuperados a partir de la frase "Self-Organizing Map"[1]. La mayor parte de los documentos publicados reportan estudios en los que se hace un uso empírico de esta técnica y existen limitados resultados teóricos que sirvan de guía para garantizar la validez de su aplicación como herramienta analítica. En consecuencia, el uso del SOM como herramienta de investigación científica requiere un entendimiento profundo del funcionamiento del algoritmo que permita el ingenio de recursos técnicos ad-hoc al problema y al escenario de aplicación.

Justificación

Una de las primeras observaciones que se ponen de relieve en este trabajo, es que la aplicación simple y directa del algoritmo SOM puede no rendir frutos e incluso producir conclusiones erróneas y que es imperativo adecuar apropiadamente: la configuración de la red neuronal, la modelación de los datos, la selección del índice de similaridad, la aplicación de métodos de normalización en el preprocesamiento y la estimación y ajuste de parámetros para lograr efectos de visualización deseados. Por esta razón, el uso experto del SOM, es considerado "un arte" por los especialistas y cada escenario de aplicación, requiere el uso de diferentes técnicas para obtener resultados relevantes.

[1] aquellos documentos indexados por Scopus en los cuales la frase aprece en el título, resumen o palabras clave.

Motivación

Motivados por esta realidad, en esta tesis se investiga la familia de algoritmos SOM que se han derivado del trabajo pionero de T. Kohonen, se analizan las peculiaridades de tres escenarios distintos de aplicación, se identifican las principales dificultades para cada peculiar escenario, y se desarrolla una serie de recursos técnicos que son de utilidad para la realización de una minería de datos relevante y confiable.

Aportaciones

Los elementos metodológicos propuestos en esta tesis se validaron y reportaron en investigaciones realizadas en coautoría con especialistas de diversos campos de aplicación. A continuación se presenta la lista de las principales publicaciones derivadas de esta investigación:

[**Villaseñor et al., 2017**] Villaseñor E.A., Arencibia R. and Carrillo H. *Multiparametric characterization of scientometric performance profiles assisted by neural networks: a study of mexican higher education institutions*. Scientometrics 101 (1), January 2017, pp. 77-104.

[**Arencibia et al., 2016**] Arencibia-Jorge, R., Villaseñor, E. A., Lozano-Díaz, I. A., and Calvet, H. C. (2016). *Elsevier´s journal metrics for the identification of a mainstream journals core: A case study on Mexico.* LIBRES, 26(1), 1-13, 2016.

[**Millán et al., 2012**] Millán V., Villaseñor E.A. Martinez de la Escalera N. and Carrillo H. *Gender differences in academic performance at UNAM*. Resources for Feminist Research; Toronto 34. 1/2 (2012): 149-164.

[**Villaseñor et al., 2012**] Villaseñor E., Guzmán M.V. and Carrillo H. *Análisis de la dinámica de la relevancia de términos MeSH y su aplicación para la visualización de traslaciones en dominios de conocimiento bio-médico.* En Memorias del Congreso Internacional de la Información. INFO'2012, La Habana, Cuba, 2012.

[**Guzmán et al., 2010**] Guzmán M.V., Carrillo H., Jiménez JL. and Villaseñor E.A. *The Art and Science of Tuberculosis Vaccines*, chapter 22: Bioinformetric Studies in TB Vaccines Research. Oxford Press, 2010 (Nueva edición 2014, chapter 3.4, pp. 474-519).

[**Villaseñor et al., 2010**] Villaseñor E.A., Guzmán M.V. and Carrillo H. *Aplicación de ViblioSOM al comportamiento temporal de los MeSH de medline*. En Memorias del Congreso Internacional de la Información. INFO'2010, La Habana, Cuba, 2010.

[**Villaseñor et al., 2008b**] Villaseñor E.A., Carrillo H., Martinez de la Escalera N. and Cruz N. *Sistemas dinámicos y visualización informétrica: Una aplicación de la red neuronal SOM*. Aportaciones Matemáticas, 2008.

[**Millán et al., 2008**] Millán V., Villaseñor E., Martinez de la Escalera N. and Carrillo H. *Género y Educación en México*. Karla Kral Ed, chapter VII: Visualización infométrica de algunas diferencias de género impresas en el egreso universitario., pages 191-210. Universidad de Colima, 2008.

[**Villaseñor et al., 2008a**] Villaseñor E.A., Carrillo H., Martinez de la Escalera N.and Millán V. *The use of weighted metric som algorithm as a visualization tool for demographic studies*. Research on Computing Science, 40:13-25, 2008.

En Villaseñor et al. [2008a], Millán et al. [2008], Millán et al. [2012] se discuten diversos aspectos de la aplicación del SOM para realizar estudios con enfoque de género. La idea central de la aplicación es utilizar al SOM como medio para visualizar las diferencias de género impresas en el egreso universitario considerando diversos factores socio-demográfico. Las visualizaciones más intuitivas y útiles son las que permiten inferir diferencias de género mediante la identificación visual de asimetrías presentes en patrones visuales (como mapas cromáticos) provistas por la red neuronal. En el trabajo se consideran tanto variables categóricas como variables numéricas. La principal aportación de esta metodología se ubica, en particular, en los problemas de análisis en donde sólo se considera una variable de control (como sexo en los estudios de género), y en los cuales es importante observar simetrías y asimetrías. En estos trabajos se muestra la capacidad explicativa de los mapas producidos.

Otro de los campos de aplicación que identificamos, en donde el análisis de la red neuronal SOM resultaba de utilidad, fue el análisis informétrico de dominios de conocimiento. En este caso los datos ("brutos") son registros bibliográficos que representan publicaciones científicas pertenecientes a un dominio de conocimiento particular. La primera publicación asociada a este trabajo se presenta en Villaseñor et al. [2008b].

En este trabajo se muestra la aplicación de índices de similitud no-métricos y técnicas de agrupamiento para la representación de estructuras jerárquicas en dominios de conocimiento, en particular el dominio de conocimiento considerado en esta aplicación es el de las matemáticas aplicadas en investigación biomédica; las estructuras representadas corresponden a las heredadas por el uso de una ontología como un vocabulario para etiquetar los registros

bibliográficos. Otro problema que se aborda en este trabajo es el de identificar agrupamientos que permiten entender la evolución de los dominios de conocimiento mediante el análisis de la aparición de tópicos. Para lograr esto se propone un índice se similitud no-métrico y métodos de clustering para el tratamiento de secuencias temporales.

En Guzmán et al. [2010] se presentan resultados que exhiben las virtudes del análisis basado en el SOM cuando se aplica al análisis temporal de la ocurrencia de términos MeSH de Medline. Posteriormente, en Villaseñor et al. [2010] se muestran las mejoras de la aplicación de estos métodos utilizando métodos no usuales de normalización y un índice de similaridad no-métrico. Un análisis más profundo de las virtudes de utilizar esta medida de similaridad en el contexto cienciométrico se encuentra en Villaseñor et al. [2012].

Finalmente, se presenta una de las ventajas de usar variantes de la red neuronal para la caracterización multifactorial. Estas aplicaciones involucraron la definición de métodos de clustering y la interpretacón de los agrupamientos resultantes, los cuales están basados en la proyección no lineal definida por el SOM. El método de análisis multimparamétrico resultó útil en la caracterización y clasificación de perfiles de productividad de distintos tipos de entidades, en particular: instituciones de educación superior Arencibia et al. [2016] y revistas Villaseñor et al. [2017].

El texto de esta tesis consta de dos partes principales: primeramente, una introducción y recuento de estado del arte, después una segunda parte donde se presentan los métodos, técnicas y las aplicaciones que nos han servido para validar el uso de las técnicas propuestas.

Índice general

Parte I

Introducción y Estado del Arte

El análisis de datos ha sufrido un cambio radical en su enfoque original durante los últimos años. El desarrollo actual de las tecnologías de la información y las capacidades computacionales redimensionan los alcances y posibilidades de la exploración de grandes bases de datos. Se pasa de la escasez de información y datos, a su proliferación; de un alto costo para el almacenamiento y procesamiento de datos, a su popularización; de la aplicación de métodos matemáticamente rigurosos de muestreo, al uso de información de poblaciones enteras.

La nueva era vislumbra: la abundancia de datos como propiedad emergente del proceso de convergencia de las TI, la proliferación de procesadores con mayor velocidad y capacidades de computo multiproceso (multicore), memorias muy rápidas y económicas y el desarrollo de hardware con la capacidad de computo masivo y paralelo (coprocesadores gráficos GPGPU). En este contexto, la aplicación de métodos analíticos basados en la inteligencia computacional para obtener nuevo conocimiento, tiene gran demanda en diferentes ámbitos.

El análisis estadístico tradicional tiene como propuesta alternativa el análisis exploratorio de datos. Este principio se utiliza en la estadística multivariada Tukey [1977]. Sin embargo, la alta dimensionalidad de la información y el gran número de registros que se generan, dificultan la aplicación de métodos estadísticos tradicionales.

El análisis exploratorio parte principalmente de la siguiente hipótesis: *mediante la aplicación de técnicas matemáticas de representación y graficación es posible develar estructuras subyacentes en los datos, las cuales aportan información valiosa respecto al dominio de apliación donde se generan los datos*. Los planteamientos sobre el análisis exploratorio de datos han evolucionado hasta lo que se conoce como, Minería de Datos (DM) y Descubrimiento de Conocimiento en Bases de Datos (KDD).

A partir de los 90, se comienzan aplicar métodos de inteligencia artificial bio-inspirados en la búsqueda de nuevo conocimiento. Las tareas de minería de datos se pueden realizar por las ventajas de las redes neuronales Bigus [1996]. Estas, se consideran como un área de conocimiento emergente que, está ganando atención en el gremio científico Berthold and Hand [1999]. En la última década se aplican métodos analíticos en grandes volúmenes de información (Big Data).

Esta nueva realidad presupone el desarrollo continuo de nuevas tecnologías y métodos para el análisis de cada vez más grades conjuntos de datos. La aplicación de métodos de inteligencia computacional es una de las tendencias de mayor proyección de desarrollo en el análisis de grandes volúmenes de datos. En particular los métodos denominados *bio-inspirados* (redes neuronales, algoritmos genéticos, autómatas celulares, entre otros), son ejemplos de la

inteligencia computacional. Hasta hace poco tiempo, estos métodos analíticos eran solamente tema de literatura científica; su aplicación a problemas del mundo real no era viable.

En muchos casos, la aplicabilidad de estos métodos sigue estando limitada a causa de su complejidad computacional, lo superlineal. Los requerimientos computacionales de su uso para el análisis de datos crece de manera desproporcionada respecto al tamaño del conjunto de datos. Por esta razón, el salto de unos cuantos miles de datos a bases de cientos o miles de millones de datos implica contar con infraestructura computacional, cuyo costo suele superar las capacidades de los grupos de investigación ordinarios. La aplicación de métodos de inteligencia computacional para el análisis de datos normalemente se encuentra restringida a la experimentación, utilizando conjuntos de datos sintéticas (no provenientes del mundo real).

Las principales razones para la poca factibilidad en la aplicación de métodos analíticos bio-inspirados en el análisis de grandes datos, era la carencia de datos y lo limitado de la infraestructura computacional disponible. Propiedades computacionales eran demostradas teóricamente y su aplicabilidad se mostraba mediante el uso de conjuntos de datos ad hoc (sin complejidad estructural y con tamaño reducido). En estos experimentos, científicos de la computación emprendieron proyectos académicos que buscaban incrementar, la precisión (o "bondad") de los métodos propuestos.

El modelo SOM ha despertado gran interés en el campo de la investigación aplicada sobre métodos analíticos basados en aprendizaje computacional para analizar datos masivos Kohonen [1982b, 1995]. Desde la perspectiva de producción científica en este dominio, es una de las *Redes Neuronales Artificiales (RNA)* más exitosas.

La red neuronal SOM ofrece ventajas como método de proyección, componente indispensable para la tarea de visualización de grandes nubes de datos multidimensionales. Este modelo de red neuronal no-supervisada es útil para el análisis y visualización. La arquitectura de esta herramienta es idónea para definir proyecciones de un espacio multidimensional hasta una retícula de neuronas bidimensionales. La principal propiedad que estos métodos aprovechan del SOM es la capacidad de definir proyecciones, de un espacio multidimensional a una retícula plana, con una alta preservación topológica.

Este trabajo se situa dentro de las aplicaciones de esta red neuronal a un tipo particular de información del mundo real: bases de datos informétricas. En los capítulos de la primera parte se introducen los conceptos fundamentales, su simbología, así como, el estado del arte del análisis y visualización informétrica.

Capítulo 1

Modelación y análisis informétrico

La Informetría es una de las disciplinas métricas en las ciencias de la información. Se define como el estudio de los aspectos cuantivativos de la información en cualquier forma, no solo como registro o bibliografía, y proveniente de cualquier grupo social, no solo científicos Tague-Sutcliffe [1992]. El término se considera integrador Egghe [2005], porque, incluye todos los estudios métricos dentro de las ciencias de la información: bibliometría (bibliografía, bibliotecas,...), cienciometría (análisis de citas, evaluación de la investigación,...), patentometría (información tecnológica, patentes,...), webometría (metrícas de la web, internet u otras redes sociales,...) y aprovecha los métodos de análisis cuantitativos desarrollados en otras disciplinas sociales como las demografía y la economía (censos, encuestas, estados financieros, registros escolares, etc).

Las disciplinas mencionadas anteriormente parten del supuesto que es posible recuperar colecciones de documentos semi-estructurados y develar el conocimiento derivado de su análisis cuantitativo. La representación semi-estructurada facilita la aplicación de métodos analíticos que buscan entre otras cosas: describir correlaciones entre variables de naturaleza diversa, reconocer patrones para clasificar los datos, estudiar cambios a lo largo del tiempo (estudios longitudinales) o representar visualmente grandes conjuntos de datos multidimensionales.

A continuación en la sección 1.1 se establecen los elementos de modelación matemática que se utilizan para procesar conjuntos de documentos semi-estructurados y definir formalmente los aspectos cuantitativos de la información. En la sección 1.2 se exponen los indicadores empleados frecuentemente en el análisis informétrico. La sección 1.3 describe alguno de los paradigmas más utilizados, para representar visualmente los conjuntos de

documentos desde la perspectiva informétrica.

1.1. Modelación Matemática de Colecciones de Documentos

Una tarea crucial que determina gran parte del procesamiento de la colección de documentos, es la manera en que sus elementos (los documentos) son matemáticamente representados. Esta modelación determina los métodos analíticos y como éstos deben ser aplicados.

En la sección 1.1.1 se presentan los diversos aspectos de la información relacionados con los documentos semi-estructurados. Los objetos que se derivan del análisis de colecciones de documentos semi-estructurados determinan objetos matemáticos multidimensionales. Los elementos cuantitativos que determinan los ojetos mencionados se presentan en la sección 1.1.2. Finalmente, en la sección 1.1.3, se define el modelo de gráfica bipartita, el cual simplifica la definición de mapeos semánticos.

1.1.1. Datos semi-estructurados

Un documento semi-estructurado se comprende de múltiples etiquetas de campos (variables) e información asociada a cada variable, que puede ser de diversa índole: indicadores numéricos, variables categóricas, marcadores temporales, descriptores semánticos, texto plano. Dada una colección de documentos semi-estructurados se puede someter a distintos tipos de procesamiento con la finalidad de obtener representaciones estructuradas de la información subyacente en los datos. Normalmente, este tipo de datos ocurre en bases de datos como: e-mails, registros bibliográficos, quejas de clientes, páginas web, diagnósticos médicos, entre otros.

Las fuentes semi-estructuradas son cada vez más comunes en distintos ámbitos del quehacer humano. Estos datos se pueden modelar como documentos semi-estructurados, utilizando formatos en lenguajes de marcado como .xml o .json. Los documentos semiestructurados pueden ser tratados de manera más eficiente por los motores de búsqueda, los cuales pueden incorporar el conocimiento de los metadatos a estructuras (tablas hash o listas ligadas). Estas se localizan en la memoria RAM y/o esquemas de programación MapReduce, para aprovechar las capacidades computacionales de los clusters de computadoras.

El análisis informétrico es el escenario de aplicación más general para los métodos planteados en esta investigación. En este caso, se considera como

punto de partida colecciones de documentos semi-estructurados. Esta característica aleja al dominio de esta tesis del terreno del procesamiento de lenguaje natural. El punto de partida son documentos semi-estructurados preprocesados, además se asume que es plausible la aplicación de operadores para el cómputo de indicadores informétricos. En ocasiones los datos son estructurados mediante el uso de metadatos semánticos (términos ordenados en ontologías). Estas representaciones permiten el acceso flexible al contenido, por tanto, las tareas de recuperación y extracción de información son más eficientes. Así mismo, se posibilita la aplicación de técnicas de análisis informétrico y minería de datos.

En este trabajo, cuando se habla de un conjunto de documentos D se está pensando en documentos semi-estructurados. Todo documento semi-estructurado $d \in D$ se puede concebir como un árbol cuyas hojas representan distintos tipos de entidades (autores (Aut), revistas (Jou), instituciones (Ins), países(Cou), referencias (Ref) , descriptores(KW)). Además, normalmente cada documento tiene una etiqueta temporal t asociada. El registro bibliográfico es un ejemplo de dato semi-estructurado de importancia en este trabajo (ver figura 1).

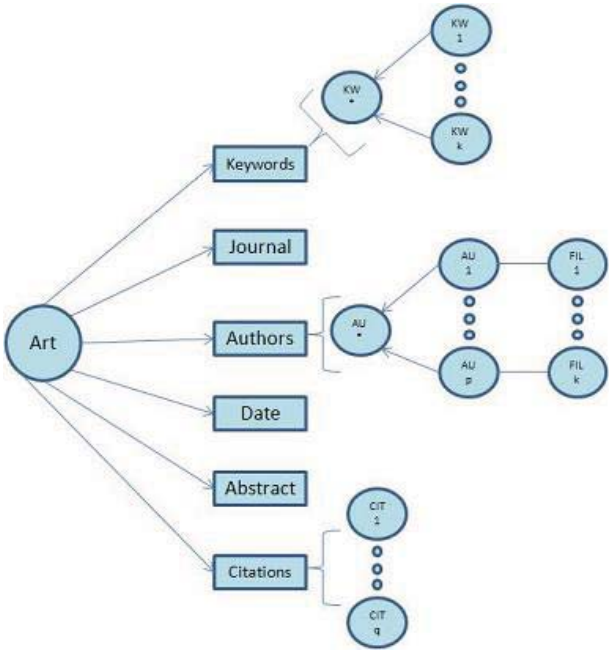


Figura 1: Modelo Semi-estructurado para un registro bibliográfico

Dado un conjunto de documentos D y un documento $d \in D$, se denotará por $KW(d)$ al grupo de descriptores asociados a d y $KW(D)$ el conjunto de

todos los descriptores asociados a cada documento en D. Esta notación se usará para todos los demás tipos de entidades que conforman al documento semi-estructurado.

1.1.2. Representación multidimensional

Tradicionalmente se emplea un conjunto de vectores de alta dimensionalidad para representar una colección D de documentos. Para construir esta representación vectorial se utiliza el conjunto de palabras clave asociadas al conjunto de documentos $KW = \{kw_1, ..., kw_n\}$.

Para cada $d \in D$ y cada $kw_i \in \mathcal{KW}$ se determina $x_i \in \mathbb{R}^n$ que, indica cual es el peso que tiene kw_i, para determinar el contenido de $x \in D$. De manera que, cada $x \in D$ queda representado por un $x \in \mathbb{R}^n$. De manera natural, la dimensión de estos vectores es el número de palabras ($n = \#KW$) que ocurren en D. Es común que, n tenga el mismo orden de magnitud que la colección de documentos $\#D = N$. Este modelo vectorial se le conoce como VSM (Vector Space Model).

Además de la representación de los documentos, un aspecto fundamental en la modelación matemática es la manera en que estos objetos son comparados. Normalmente, es parte de una medida de disimilaridad $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$. Es posible utilizar alguna función de distancia sobre $\mathbb{R}^n$ como espacio métrico. Sin embargo, la práctica demuestra que, la métrica del coseno del ángulo es más adecuada para medir la similitud semántica entre vectores $X \subset \mathbb{R}^n$ obtenidos por la aplicación del VSM. En este caso, el espacio topológico más adecuado para representar el universo $\mathbb{X}$ de todos los posibles documentos es la hiperesfera unitaria $\mathbb{S}^n$.

En la mayoría de estas aplicaciones del VSM, el uso de un método de reducción de dimensionalidad es crucial para el procesamiento de recopilaciones masivas de documentos. Muchos de estos métodos confrontan la problemática típica de minería de textos donde los documentos son no-estructurados. En nuestro caso, los documentos están etiquetados por expertos, por tanto, la identificación y organización jerárquica de las palabras clave es a través de una ontología.

Este trabajo parte del siguiente modelo general: con el preprocesamiento de colecciones de documentos semi-estructurados, todo documento semi-estructurado $d \in D$ puede ser modelado como un arreglo multidemensional

$$d \in \mathbb{R}^n \times \mathbb{S}^m \times \mathbb{Z}_2^k \times \mathbb{T}. \tag{1.1}$$

Donde las componentes de $\mathbb{R}^n$ corresponden a mediciones definidas como indicadores informétricos relacionados al documento (ver sección 1.2), $\mathbb{S}^m$ es un

espacio semántico multidimensional que caracteriza al registro, $\mathbb{Z}_2^k$ representan variables categóricas que a su vez también pueden representar descriptores (semánticos) y $\mathbb{T}$ es la variable temporal, que en el caso de los documentos se relaciona con la fecha de publicación.

Dada esta representación la dimensión dim del documento es

$$dim = n + m + k + 1.$$

Por lo general, n es relativamente pequeño, m es muy grande pero se puede reducir, y k no es tan grande pero puede ser irreducible:

$$n < k << m.$$

1.1.3. Modelos gráficos de documentos

El modelo multidimensional definido en la sección (1.1) es muy general; sin embargo, solo se restringe a modelar los objetos y no las relaciones entre estos. En ocasiones resulta útil contar con una descripción sintáctica de los documentos.

EL concepto de gráfica es muy últil para modelar matemáticamente las relaciones entre un conjunto de objetos. Una gráfica se define como una dupla $(\mathcal{V}, \mathcal{A})$, $\mathcal{V}$ es le conjunto de vértices, que en nuestro caso serán documentos y $\mathcal{A}$ es el conjunto de aristas. El conjunto de aristas $\mathcal{A}$ queda definido por una relación $\mathcal{A} : \mathcal{V} \to \mathcal{V}$.

Una situación común en los registros bibliográficos de documentación científica, es contar con un conjunto jerárquicamente ordenado de palabras clave (ontología). Las cuales son útiles para etiquetar y caracterizar semánticamente documentos dentro de un dominio específico. En este caso se parte de un conjunto de registros bibliográficos D, etiquetados por un conjunto de descriptores KW[1].

Dado el conjunto de documentos D y el conjunto de palabras clave $KW(D)$ que ocurren en D, hay una gráfica simétrica bipartita asociada $\mathcal{B}(\mathcal{V}, \mathcal{E})$ con un conjunto de nodos $\mathcal{V}$ y un conjunto de aristas $\mathcal{E}$ dados por:

$$\begin{aligned} \mathcal{V} &= D \cup KW \\ \mathcal{E} &= \{(d, kw) \mid kw \in KW(d)\}. \end{aligned} \tag{1.2}$$

La representación de documentos utilizando este modelo se vuelve más simple e intuitiva. Además, son más eficientes en cuanto al uso de memoria en

[1]Esta suposición no es restrictiva. Actualmente, cada vez más organizaciones cuentan con sus propias bases de conocimiento mediante el uso de técnicas de *ingenierá del conocimiento* utilizando tecnologías semánticas.

comparación con la representación tradicional del vector de alta dimensionalidad.

Este modelo de representación de documentos se utiliza para el agrupamiento simultáneo de las palabras en $KW(D)$ y los documentos en D utilizando un algoritmo de clustering espectral Dhillon [2001]. Una de las premisas que hay detrás de este algoritmo es la dualidad entre agrupamientos de palabras y documentos, lo que significa: *el agrupamiento de palabras induce el agrupamiento de documentos, en tanto que el agrupamiento de documentos induce el agrupamiento de palabras.*

En el capítulo 6 se presenta un enfoque similar al presentado en Dhillon [2001] Pessiot et al. [2010]. Consideramos la gráfica bipartita (1.2) como un modelo para un conjunto de documentos y analizamos el agrupamiento de documentos inducido por el agrupamiento de palabras, considerando tanto la dualidad entre agrupamientos de palabras y documentos como las premisas contextuales. En nuestro caso, el contexto estará dado en términos de espacio (dominio de conocimiento) y tiempo, esto mediante el análisis de las secuencias temporales de la ocurrencia de palabras.

1.1.4. Espacios de Similitud

Uno de los conceptos clave en la configuración de métodos para el análisis y visualización de colecciones de documentos es el de función de similitud (o disimilitud) entre los documentos.

Un índice de similitud para $\mathbb{X}$ es una función $s : \mathbb{X} \times \mathbb{X} \to [s_{\text{mín}}, s_{\text{máx}}] \subset \mathbb{R}$ donde $s_{\text{mín}}$ y $s_{\text{máx}}$ son la mínima y la máxima similitud. Esta función debe cumplir las siguiente propiedades:

- $s(x, x) = s_{\text{máx}}$ para toda $x \in \mathbb{X}$ (reflexividad).
- $s(x, y) = s(y, x)$ para toda $x, y \in \mathbb{X}$ (simetría).

Dado un universo muestral $\mathbb{X}$ y una función de similitud $s : \mathbb{X} \times \mathbb{X} \to [s_{\text{mín}}, s_{\text{máx}}]$ a la dupla $(\mathbb{X}, s)$ se le conoce como **Espacio de Similitud**.

Dados $x, z \in \mathbb{X}$ y un índice de similitud $s : \mathbb{X} \times \mathbb{X} \to [s_{\text{mín}}, s_{\text{máx}}]$, $s(x, y)$ indica que tan similares son x y z. Cabe señalar que lo apropiado de un índice de similitud para un conjunto de objetos dado, radica principalmente en que tan apegado es éste índice a la intuición o sentido común. Es decir, aunque se puedan definir muchas maneras de medir la similitud y todas ellas arrojen resultados distintos, la mejor de ellas será la que arroje resultados que sean más fáciles de interpretar en términos del dominio específico de aplicación.

En el caso donde $\mathbb{X}$ es un espacio métrico, se suele interpretar la distancia como medida de disimilitud. De manera que a menor distancia, mayor simi-

litud. Recordemos que una función de distancia d sobre un espacio métrico $\mathbb{X}$ es de la forma $d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}^+$ [2] tal que:

- $d(x, y) = 0 \Leftrightarrow x = y$ (reflexividad)
- $d(x, y) = d(y, x)$ (simetría)
- $d(x, z) \leq d(x, y) + d(y, z)$ (desigualdad del triángulo)

para cualesquiera $x, y, z \in \mathbb{U}$. La siguiente proposición establece la manera de crear índices de similitud a partir de funciones de distacia.

Proposición 1 *Sea $d(x, y)$ una función de distancia y sea $f : \mathbb{R}^+ \to [s_{\text{mín}}, s_{\text{máx}}]$ una función monótona decreciente tal que $f(0) = s_{\text{mín}}$ y $\lim_{z\to\infty} f(z) = s_{\text{máx}}$ entonces $f(d(x, y) : X \times X \to [s_{\text{mín}}, s_{\text{máx}}]$ es un índice de similitud.*

Una generalización del concepto de distancia es el de disimilitud. Esto es, una función $d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ es de disimilitud si exite un índice de similitud s y una función f como la de la proposición 1 tal que

$$s = f \circ d. \tag{1.3}$$

Una observación interesante es que a pesar de que cualquier medida de distancia define una familia no numerable de índices de similitud, no todo índice de similitud proviene de una función de distancia. Lo anterior es consecuencia de que la desigualdad del triángulo es una restricción no necesaria para las funciones de disimilitud. A las funciones de disimilitud que no son funciones de distancia les llamaremos funciones de disimilitud no métricas. En la sección 6.2.3 se define una función de similitud no métrica adecuada para medir similitud entre secuencias temporales de ocurrencias libres de escala. Esta función de similitud será de utilidad para definir un mapeo temporal de las colecciones de documentos basado en el algoritmo SOM. Este mapeo temporal constituye una de las principales aportaciones de esta tesis. La definición de la función de similitud y aplicación para el análsis de la evoulución de dominios de conocimiento se muestra en el capítulo 6.

1.2. Elementos de Análisis Informétrico

La Bibliometría se considera una de las primeras técnicas cuantitativas utilizadas históricamente para el análisis de información (datos semiestructurado). Su gran aceptación en la comunidad de analistas radica en la

[2] $\mathbb{R}^+ = \{x \in \mathbb{R} | x \geq 0\}$

definición de diversos indicadores basados principalmente en conteos a partir de los registros bibliográficos donde se reportan resultados de la actividad científica.

La paternidad del término Bibliometría se le atribuye a Alan Pritchard, quien en 1969 publicó el ensayo "Bibliografía Estadística o Bibliometría?". En el documento se definó el término Bibliometría como alternativa de la Bibliografía Estadística. Este último podía traer confusión con Estadística o una Bibliografía sobre Estadística. El enfoque de Pritchard se centra en la definición de indicadores para describir algún aspecto relevante en el proceso de comunicación de la información escrita.

En el mismo año de 1969, los rusos Nalimov y Mulchenco acuñan el término Naukometriya para designar los estudios científicos realizados a la actividad de la investigación científica. Esta disciplina busca identificar patrones, estructuras y tendencias en la evolución de la ciencia como un sistema productivo. Considerando el sistema social conformado por los individuos (científicos) y las instituciones académicas. Posteriormente, en occidente a la Naukometriya se le conocerá como Scientometrics (o Cienciometría en Español).

En general, las disciplinas métricas parten de colecciones de documentos en forma de datos semiestructurados. Por tanto, el planteamiento de la modelación de datos primarios de la sección anterior aplica a todas estas disciplinas. En este trabajo nos centraremos principalmente en los indicadores cienciométricos. Sin embargo,los elementos intrínsecos en estas definiciones pueden ser fácilmente extendidos a las otras disciplinas métricas.

1.2.1. Indicadores de producción

Sea $D \subset \mathcal{D}$ y una entidad $x \in \mathcal{E}$ la producción de x se define como

$$Pr_D(x) = \{d \in D \mid x \in d\} \tag{1.4}$$

$$pr_D(x) = \sharp Pr(x).^{3} \tag{1.5}$$

Las definiciones 1.4 y 1.5 se pueden extender a conjuntos de entidades. Dado un conjunto de entidades $E \subset \mathcal{E}$, la producción de E es simplemente

$$Pr_D(E) = \bigcup_{x \in E} Pr_D(x)$$

$$pr_D(E) = \sharp Pr_D(E)$$

[3] $\sharp$ se utilizará para denotar cardinalidad

Indicadores de Producción

En las mediciones para evaluar la producción de una entidad $x \in E$, se considera un intervalo de tiempo T compuesto por un conjunto de años. Considerando que, para cada $t \in T$ se pueden considerar D_t los documentos producidos en el año t. En particular, se pueden considerar los documentos publicados en revistas indexadas por *Web of Science* y *Scopus*, D_t^{WoS} y D_t^{Scopus}. De manera que se computan los indicadores de producción para cada conjunto de documentos.

Simbología	Fórmula	Nombre
$ANdoc(x)$	$pr_{D_t}(x)$	*Producción anual de documentos*
$ANdoc^{WoS}(x)$	$pr_{D_t^{WoS}}(x)$	Producción anual de documentos en *Web of Science*
$ANdoc^{Scopus}(x)$	$pr_{D_t^{Scopus}}(x)$	Producción anual de documentos en *Scopus*

En la gráfica de la figura 2 se puede apreciar la evolución de los indicadores de prducción donde x es el país México y $T = \{1996, ..., 2011\}$. Además, en esta misma gráfica se muestra la evolución del indicador $ANnrs$ que se definen como el número de miembros del SNI (Sistema Nacional de Investigadores) en cada año.

De la gráfica se puede concluir que, existe un crecimiento paralelo entre la producción y el número de investigadores en el SNI. Sin embargo, no queda explícito, que tan eficiente es México produciendo investigación científica, para esto último es necesario definir indicadores de productividad.

Indicadores de productividad nacional

La productividad la entendemos como eficiencia en la producción por lo que para medirla es necesario considerar los recursos involucrados en los procesos productivos. En este caso, los investigadores constituyen el capital humano. A continuación se definen algunos indicadores de productividad nacional para el caso mexicano, considerando el número de investigadores en el SNI ($Nnrs$).

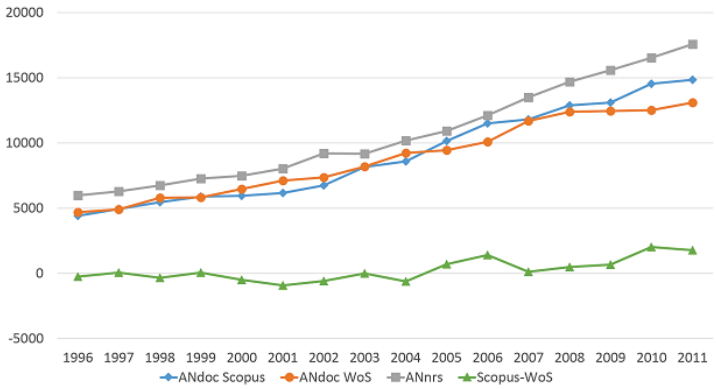


Figura 2: Crecimiento de la produción científica mexicana

Simbología	Fórmula	Nombre
$NSP(x)$	$\frac{ANdoc(x)}{Nnrs}$	*Productividad Científica Nacional*
$NSP^{WoS}(x)$	$\frac{ANdoc(x)^{WoS}}{Nnrs}$	*Productividad Científica Nacional en WoS*
$NSP^{Scopus}(x)$	$\frac{ANdoc^{Scopus}(x)}{Nnrs}$	*Productividad Científica Nacional en Scopus*

Producción y productividad institucional

A continuación se definen indicadores de producción y desempeño para instituciones. En este caso, y es una institución dentro del país x.Para evaluar el desempeño en la producción de una institución resulta útil considerar un período T relativamente pequeño de tiempo (cinco años) y calcular el promedio de los indicadores calculados en cada año. Además, será necesario considerar el número de miembros del SNI para cada institución en cada año, $NRS(y,t)$.

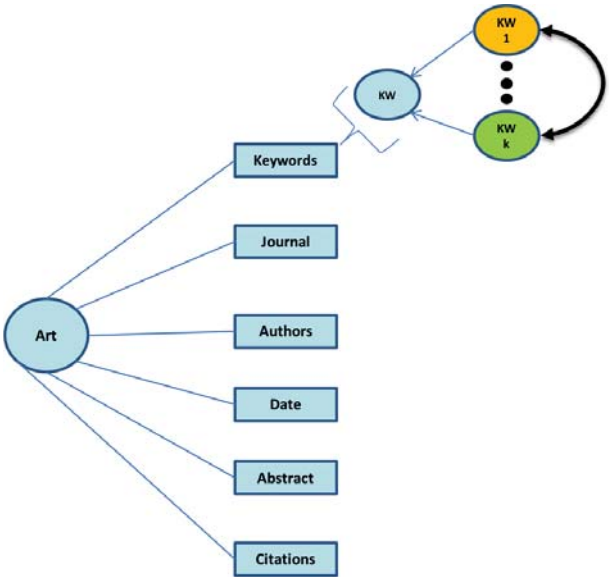


Figura 3: Relaciones entre entidades que se establecen a partir del cómputo de co-ocurrencias

Simbología	Fórmula	Nombre
$AINdoc(y)$	$\frac{1}{\#T}\sum_{t\in T} pr_{D_t}(y)$	*Producción institucional anual*
$INnrs(y)$	$\frac{1}{\#T}\sum_{t\in T} NRS(y,t)$	*Promedio de SNIs por institución*
$ISP(y)$	$\frac{AINdoc}{INnrs}$	*Productividad Científica institucional*

1.2.2. Indicadores relacionales

Los indicadores relacionales miden la relación entre dos entidades. Estos indicadores se basan en el concepto de coocurrencia. Dadas dos entidades a y b, la coocurrencia de ellas $coo(a, b)$ en D,

$$coo(a, b) = \#\{d \in D \mid a, b \subseteq d\}$$

es el número de documentos en donde las dos entidades aparecen.

Dados dos conjuntos de entidades $A = a_i$ y $B = b_j$ la matriz de co-ocurrencia de A con B se define como:

$$COO(A, B)_{i,j} = coo(a_i, b_j) \tag{1.6}$$

cuando $A = B$ la matriz de co-ocurrencia se denotará por $COO(A)$. En este caso la red resulta simétrica en los valores de su diagonal corresponden a los valores $\{pr(a_i)\}_{i=1:n}$ con $pr()$ definido como en (1.5).

Las matrices de co-ocurrencia definen redes susceptibles de análisis estuctural. Dado el gran tamaño de estas matrices, una práctica común en estudios métricos es, descartar entidades de baja frecuencia utilizando un umbral arbitrario. Por tanto, el análisis se enfoca en un subconjunto más pequeño de la base de datos y no se considera una gran cantidad de información.

Supongamos que, se tiene un conjunto de documentos D, descrito por un conjunto de palabras clave KW. Utilizando (1.6) se puede construir $COO(KW)$. Esta matriz define la *matriz de adyacencia pesada* de una *red de palabras clave* $\mathcal{KW}_D$. Cuando el conjunto de documentos D, es tal que $\#D = N$ es grande, el tamaño n del conjunto de palabras clave KW suele ser también grande. Dada la alta dimensionalidad de la matriz $COO(KW)$, su análisis se ha enfocado en métodos para reducción de dimensión o métodos para graficar las redes de palabras inducidas por la relación de coocurrencia $\mathcal{KW}_D^{COO}$ para develar su estructura.

En el capítulo 6 de esta tesis se muestran los resultados de la graficación de redes de co-ocurrencia utilizando un conjunto D que está asociado con un dominio específico de investigación (en el área biomédica). Estos conjuntos de documentos tienen la propiedad de estar semánticamente caracterizados por un conjunto de descriptores KW.

1.2.3. Indicadores de Impacto

La medida de impacto más simple se define a nivel de cada documento. Esto se realiza a través de una red de citas, la cual se define a partir de una relación asimétrica entre documentos. De manera que, dados $d, \hat{d} \in D$ la relación de citación $\rightarrow_{cit}$ se define como:

$$d \rightarrow \hat{d} \Leftrightarrow_{cit} \hat{d} \in Ref(d).$$

Se define la gráfica $\mathcal{C} = (D, \rightarrow_{cit})$ que resulta de gran utilidad para medir el impacto que cada documento tiene en el desarrollo de la investigación científica. El impacto de cada $d \in D$ es el grado de centralidad del documento en esta red,

$$imp_D(d) = deg_{\mathcal{C}}(d).$$

A partir de esta medida de impacto para los documentos se pueden definir medidas de impacto a nivel de las instituciones. Por ejemplo, se puede

considerar el impacto de la producción institucional como:

$$imp_D(y) = \frac{\sum_{d \in Pr(y)} imp_D(d)}{pr(y)}.$$

Sin embargo, imp_D no es útil para medir el impacto de la producción científica de una institución comparada con otras instituciones o con el impacto que en promedio tienen los artículos científicos a nivel internacional. Para esto resulta útil considerar el número de citas que en promedio reciben los artículos,

$$imp_D^{World} = \frac{\sum_{d \in D} imp_D(d)}{\#D}$$

Además, los conjuntos de referencia, $Q_1^D \subset Rev(D)$ primer cuartil de las revistas más citadas y $Dec_1^D \subset D$ es el primer decil de los documentos más citados en Dz.

Otro aspecto importante en el análisis cienciométrio es la segmentación de los conjuntos de documentos por *áreas de conocimiento*. En la práctica se ha observado que, cada disciplina tiene su propio patrón de citas. Se asume que, el conjunto de documentos D puede ser segmentado considerando un conjunto de áreas de conocimiento $\{KD_1, ..., KD_q\}$ es a su vez una partición de D. Dada una institución y se definen los siguientes indicadores de impacto.

Simbología	Fórmula	Nombre
$\%Q_1(y)$	$\frac{\#\{d \in Pr(y) \mid Rev(d) \in Q_1^D\}}{pr(y)}$	*Porcentaje de artículos en revistas del primer cuartil*
$\%Exc(y)$	$\sum_{i=1}^{q} \frac{\#\{d \in Pr_{KD_i}(y) \mid d \in Dec_1^{KD_i}\}}{q * pr_{KD_i}(y)}$	*Porcentaje de excelencia*, que es medido como el promedio de las porciones de artículos dentro del primer decil $Dec_1^{KD_i}$ en cada dominio de conocimiento.
$NI(y)$	$\sum_{t \in T} \frac{1}{\#T} \sum_{i=1}^{q} \frac{imp_{KD_i^t}(y)}{imp_{KD_i^t}^{World}}$	*Impacto normalizado*

En el capítulo 4 se propone un método basado en redes SOM para caracterizar los perfiles de desempeño multiparamétricos. Estos perfiles se construyen mediante el cómputo de diversos indicadores de desempeño, lo cuales

se construyen a partir de estos dos indicadores elementales. A continuación se exponen los elementos matemáticos básicos en la composición de estos indicadores con la finalidad de entender la naturaleza de los datos. Se discutirá cada aplicación para poder dimensionar adecuadamente las aportaciones metodológicas de este trabajo.

1.2.4. Panel Informétrico

La información sobre la dinámica de un dominio de conocimiento se puede utilizar para analizar la evolución de dicho dominio y ayudar a las investigaciones a responder preguntas tales como: ¿Cuáles son los tópicos que son el objeto principal de la investigación durante un periodo en particular? ¿Cómo cambia con el tiempo el foco temático de la comunidad investigadora?

Existen distintos esfuerzos que buscan responder este tipo de preguntas para esto aprovechan las técnicas de minería de datos temporal. Por ejemplo, en Mei and Zhai [2005] se presenta un método para el descubrimiento de patrones de evolución de temas en los temas presentes en conjuntos de documentos. En este caso se parte de una secuenencia de documentos $\{d_t\}_{t\in T}$ donde T es un conjunto discreto de etiquetas temporales y d_t con fecha de publicación t.

En el caso de dominios de conocimiento sabemos que muchos documentos son publicados en la misma fecha. Por ejemplo, todos los artículos de un mismo número de una revista especializada tendrán asociados la misma fecha de publicación. Además, en el análisis de evolución de los temas normalmente se utilizan como unidad temporal los años.

Recientemente se muestran avances en el análisis de la evolución en los temas de investigación en el campo *Information Retrieval*. En Chen et al. [2017] se muestra un método basado en modelos generativos, en el cual el conjunto de documentos se divide en una partición $\{D_{T_1}, ..., D_{T_S}\}$ de manera que cada T_i corresponde a un periodo de años preestablecido y cada D_{T_i} corresponde al conjunto de documentos publicados durante T_i.

En este trabajo consideramos una manera distinta de estudiar esta dinámica de dominios de conocimiento. El planteamiento consiste en partir de la sucesión de conjuntos de documentos semi-estructurados de la forma

$$\{D_t\}_{t\in T}, \tag{1.7}$$

donde cada D_t es un conjunto de documentos semi-estructurados etiquetados con t y posteriormente aplicar métodos de clustering temporal obtener los subperiodos de análisis. A la estructura de información definida en (1.7) la denominaremos *Panel Informétrico*.

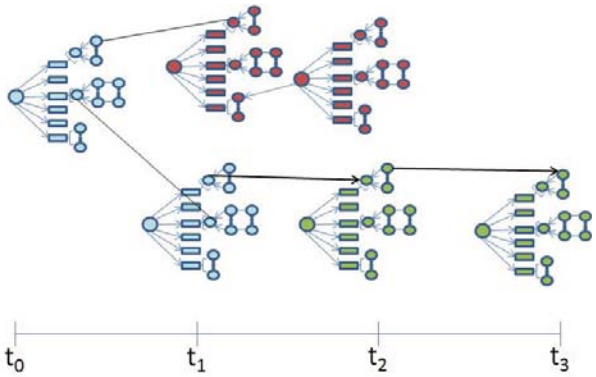


Figura 4: Evolución de las relaciones entre entidades para establecer medidas de similaridad entre datos con estructura compleja de los registros bibliográficos

La partición temporal del conjunto de documentos $\{D_t\}_{t \in T}$ permite realizar mediciones en distintos momentos, de manera que se puedan establecer tendencias o patrones temporales que acusen a distintos aspectos de la evolución de los dominios de conocimiento.

1.3. Visualización de Información

En esta tesis presentamos algunas aplicaciones de redes neuronales para el procesamiento, análisis y visualización de colecciones de documentos semi-estructurados y con etiqueta temporal. En esta sección referimos varios esquemas y recursos de visualización que son útiles para representar el conocimiento obtenido a partir del análisis neurocomputacional. Cada uno de ellos ofrece ventajas y capacidades que son complementarias entre sí. En el capítulo 6 mostraremos, en varios escenarios de aplicación, como pueden usarse las redes neuronales para generar cada formas de visualziación.

Los métodos basados en redes SOM se presentarán con mayor detalle en la sección 3.2. A continuación se exponen algunos de los problemas de visualización de información de colecciones de documentos que pueden ser abordados con el SOM.

El uso de métodos adecuados para la visualización de grandes conjunto de datos tiene una importancia estratégica para los métodos de análisis inteligente de datos. La visualización de información puede ser entendida como un proceso asistido por la computadora, en el cual se busca develar características de un fenómeno abstracto al transformar datos en formas visuales Brachman and Anand [1996]. Este campo estudia como representar grandes volúmenes de información no numérica como: texto, código de software, redes

sociales, etc..

Figura 5: La intención de la visualización de información es optimizar el uso de nuestra percepción y la habilidad de nuestro pensamiento para tratar con fenómenos que por si solos no pueden ser representados visualmente en un espacio bidimensional Bertin [1999].

Desde la perspectiva del KDD, la integración de herramientas de análisis automático y exploratorio, con la ayuda de técnicas de visualización, permitirá un descubrimiento de conocimiento más eficiente y efectivo Keim et al. [2010].

Los problemas generales que hoy en día afronta la visualización de información son Chen [2005]:

1. Crear representaciones visuales con mayor resolución espacial evitando las "nubes" de datos o etiquetas de datos.

2. La cantidad de datos procesados no sean un límite para la percepción.

3. Integrar la visualización con otros recursos multimedia como la voz.

4. Interfaces visuales que permitan la interacción temporal con el usuario.

Hasta el momento los adelantos en función de la solución de estos problemas están a nivel de laboratorio y necesitan del apoyo de una gran capacidad computacional. Con respecto a la aplicación de estas técnicas en el contexto de estudio informétricos, se conocen algunos progresos como la utilización de iconos para representar conceptos (redes de citas, autores, revistas, etc.) White et al. [2001], la confección de gráficos con presentaciones jerárquicas, la elaboración de mapas que representan estructuras de conglomerados, efectos de *zoom* para mostrar detalles, animación y perspectiva en tres dimensiones. Chen [2006a].

Los avances en la visualización de información ofrecen herramientas prometedoras para las estructuras del conocimiento y su desarrollo de una manera cada vez más intuitiva. Los métodos más empleados para el mapeo de

dominios de conocimiento se basan en el análisis de las redes de citas y co-citas. No obstante, los algoritmos basados en la teoría de gráficas estándar no logran representar de manera intuitiva la estructura de la red de citas debido a la presencia de un gran componente bien conectado.

Existen sistemas que implementan diferentes métodos para obtener representaciones visuales a partir del procesamiento de la información contenida en grandes colecciones de documentos. Estos consideran diversas representaciones de documentos y estructuras asociadas. Los métodos de visualización que utilizan son variados y no existe uno que predomine sobre los demás. Los métodos basados en redes neuronales SOM han ganado una gran popularidad por las ventajas computacionales que ofrecen para la visualización y el buen desempeño en comparación con otros métodos computacionalmente más demandantes.

Actualmente existen varios sistemas que implementan distintos métodos para obtener representaciones visuales a partir del procesamiento de la información contenida en grandes colecciones de documentos, los cuales consideran diversas representaciones de documentos y estructuras asociadas a los mismos. Los métodos de visualización que utilizan son variados y no existe aún uno que predomine sobre los demás. Sin embargo, los métodos basados en redes neuronales SOM han ganado una gran popularidad por las ventajas computacionales, las ventajas que ofrecen para la visualización y el buen desempeño en comparación con otros métodos computacionalmente más demandantes.

1.3.1. Proyección de etiquetas

Los *tag cloud* son representactaciones visuales de colecciones de documentos en forma de nubes de palabras. Los *tag* se refieren a las palaberas o términos más importantes que ocurren en la colección de documentos. El tamaño con el que se despliegan estas palabras hace referencia a la relevancia que este *tag* tiene dentro de la colección. Esta técnica es muy utilizada sobre todo para representar la información contenida en sitios web.

La interface PubCloud despliega un *Tag Cloud* a partir de un conjunto de resúmenes recuperados de la base de registros bibliográficos PubMed. Esta interfaz tiene la ventaja de presentar la información de manera descriptiva y reducir la frustración, sin embargo, es poco efectiva en revelar relaciones entre los términos Kuo et al. [2007].

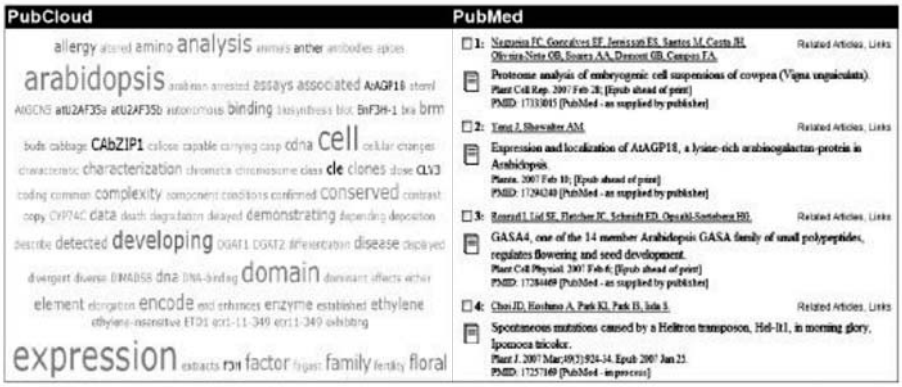


Figura 6: Interface PubCloud (tomada de Kuo et al. [2007])

1.3.2. Visualización de Conglomerados de Documentos

Una vez definida una proyección de los documentos, es posible proyectar *clusters* de documentos usando los métodos de *clustering* como los particionales o los jerárquicos. Estas proyecciones son útiles para representar la densidad de los objetos similares en forma de *cluster (clustering)*.

La interpretación de los agrupamientos de documentos es uno de los problemas fundamentales de los métodos de visualización de clustering. Por lo general, no todas las personas explican de igual forma los mapas.

Uno de los paradigmas más interesantes y con mayores adeptos en la representación visual de colecciones de documentos es el denominado 'metáfora del paisaje'. Este método busca representar las colecciones de documentos dentro de mapas topográficos que representan a los 'espacios documentales'. En estos mapas los documentos son representados por puntos y la cercanía entre dos documentos acusa su similaridad.

Estos métodos dividen el espacio por áreas temáticas de manera que los documentos son proyectados a aquella área temática a la que pertenecen. Las áreas temáticas definen territorios similares a los países y continentes en un mapa geográfico. En muchos casos se utiliza el relieve del terreno para resaltar la importancia de un tema (ver figuras 7 y 8).

La mayoría de estos métodos utilizan un modelo de documentos, basado en una representación vectorial, donde cada entrada del vector está determinado por la ocurrencia de una palabra o término en el documento. De manera que, la similitud entre los documentos está dada en términos de alguna función de distancia entre las representaciones vectoriales. Este tipo de representación ofrece problemas principalmente ligados con la alta dimensionalidad de los vectores.

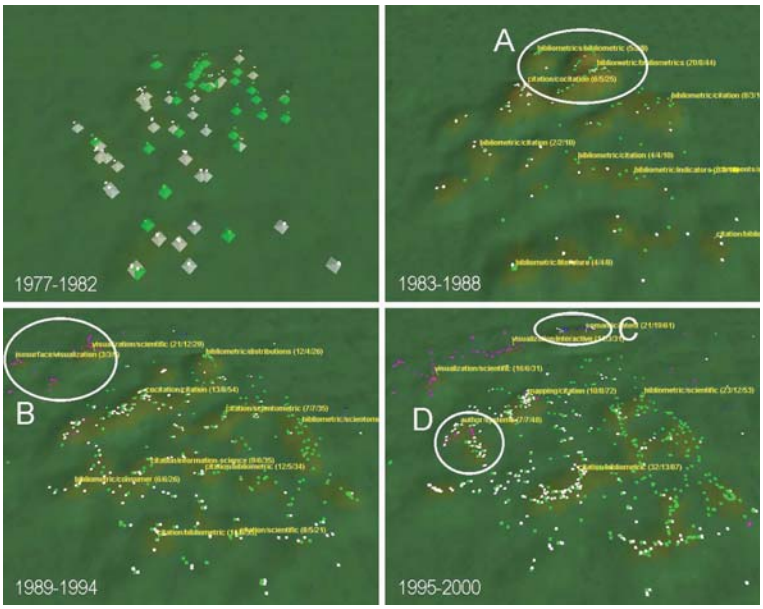


Figura 7: Sistema VxInsight (tomado de Börner et al. [2003])

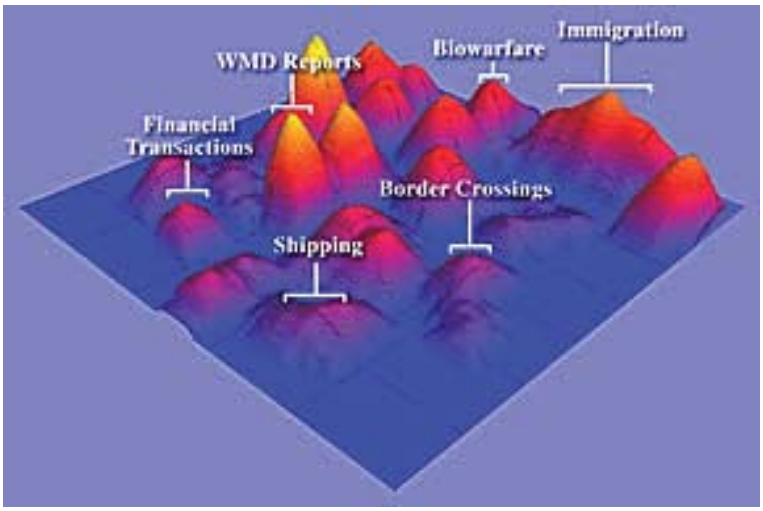


Figura 8: Interface ThemeView desarrollado por el *Pacific Northwest National Laboratory.*

Al respecto Ed Noyons señala que, una de las ventajas de los mapas, desde un enfoque métrico (ya sea sobre artículos científicos o patentes), es la posibilidad de hacer representaciones de la combinación de cualquiera de los elementos presentes en los campos del registro bibliográfico correspondiente Noyons [2001]. Estas representaciones se basan en una medida de relación entre dichos campos, relacionada con la co-ocurrencia de determinados elementos en los registros de la base de datos. De esta forma, los elementos que aparecen juntos están relacionados y la fortaleza de esa relación se pone de manifiesto en la frecuencia de la co-ocurrencia.

1.3.3. Visualización de Redes de Documentos

Estos métodos representan la colección de documentos como una red. La red se define a partir de relaciones entre documentos; como son, de citas, co-citas, acoplamiento bibliográfico o de autores y similitud semántica. Los métodos de graficación tienen básicamente dos componentes: proyección de nodos y podado de aristas. Los métodos de poda más utilizados son Path Finder y MST (Minimum Spaning Tree).

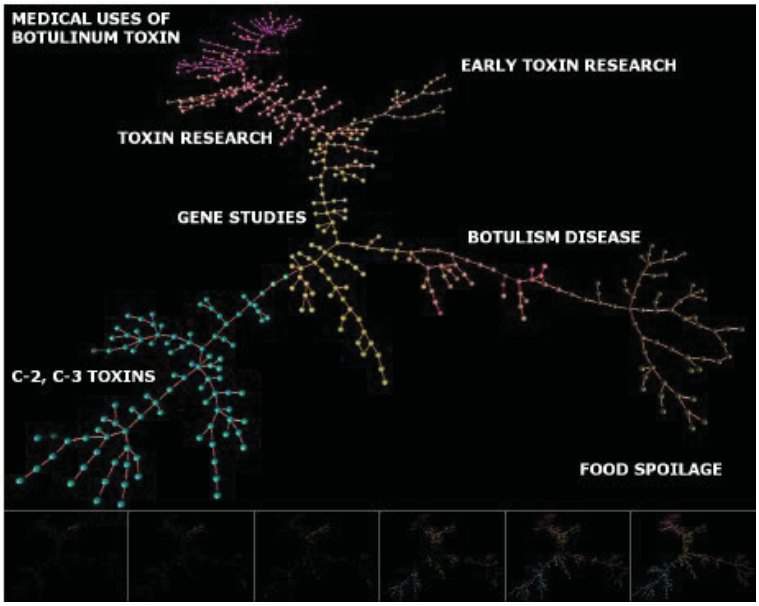


Figura 9: Sistema Citespace Chen [2006b], aplicación del método Path FInder

1.3.4. Limitaciones

Los analistas proponen inferencias con un nivel de incertidumbre porque, dependen de la apreciación de las variables representadas. A nuestro juicio, esta es una de las limitantes del mapeo métrico y de muchas de las visualizaciones. Al respecto Katty Börner explica: "recomiendo altamente probar la visualización con usuarios reales. Sorprende ver lo poco que los usuarios entienden sobre la visualización, lo difícil que es para ellos usarla y como intentan abusar de ella. Por tal motivo, resulta altamente beneficioso involucrar a los usuarios en el proceso de diseño desde el primer día"(2005).

Los problemas en los métodos de visualización de colecciones de documentos están relacionados con el diseño y arquitectura de la información. En

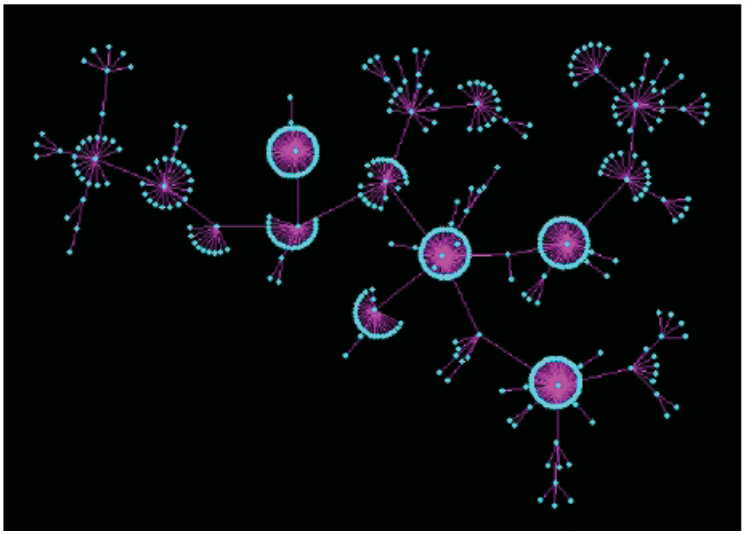


Figura 10: Método MST implementado en Chen [2006b].

particular, se listan algunos para el caso de la aplicación de estos métodos al análsis informétrico Guzmán [2010]:

- Varias de las interfaces visuales aún no logran el propósito fundamental de la visualización: que el usuario pueda apreciar de forma rápida y sencilla los patrones relevantes.
- Muchos métodos arrojan como resultado conglomerados de etiquetas o nombres de variables que son mostradas como una nube de puntos de difícil interpretación.
- Los métodos computacionales involucrados son muy costosos en tiempo y espacio, esto limita notablemente el tamaño de la colección en función de las capacidades de cómputo.
- Se necesitan mejorar aspectos tales como el uso del color, posición de las variables y hacer más evidente sus asociaciones, área visual óptima, etc.
- Se necesita facilitar la interactividad entre el usuario y la representación visual.

Capítulo 2

Algoritmos SOM

Las redes neuronales constituyen un recurso muy valioso para la minería de datos. Entre las arquitecturas neuronales más usadas se encuentran los perceptrones multicapa y las redes neuronales auto-organizadas entrenadas por medio de algoritmos SOM (Self-Organizing Maps). En esta tesis se estudian las redes SOM, así como, se desarrollan métodos y técnicas para poder aplicarlas en diferentes escenarios.

Las redes SOM fueron presentadas por T. Kohonen en 1982 donde introduce el *Basic SOM* Kohonen [1982a]. Desde entonces, aumentó la producción de artículos donde aplican el SOM para analizar bases de datos de muy variada naturaleza. A partir de la red neuronal original de Kohonen, se ha derivado una serie de algoritmos de entrenamiento que constituyen lo que se conoce como la referida familia de algoritmos SOM.

Una característica de esta red neuronal es que no requiere un entrenamiento supervisado, lo cual la hace útil para descubrir la estructura y el conocimiento contenido en una base de datos. Los algoritmos SOM, no sólo son útiles para analizar los datos, sino también, para producir espectaculares visualizaciones, que revelan la estructura y los patrones de organización del conjunto de datos. Este proceso se realiza automáticamente, y los escenarios de visualización producidos son llamados mapas auto-organizados.

Estas visualizaciones se generan gracias a la peculiar arquitectura de esta red neuronal, la cual consta de dos capas, una de entrada y otra de salida. La capa de entrada recibe las señales (datos) y los proyecta a la capa de salida conformada por un arreglo reticular bidimensional de celdas.

Generalmente, los datos son modelados matemáticamente como vectores pertenecientes a un espacio multidimensional. A través del entrenamiento de la red, adaptativamente se ajustan los pesos de las conexiones sinápticas y se determina una proyección no lineal del espacio multidimensional al plano que contiene la capa de salida de la red neuronal.

El entrenamiento implica la presentación recurrente del conjunto de datos a la red neuronal que se analiza. Los algoritmos de entrenamiento usan operadores de competencia entre los datos, que dan como resultado una neurona ganadora en la capa de salida. Así, cada neurona de la capa de salida se convierte en una celda receptora del conjunto de datos ganados por ella mediante el proceso de competencia.

En la capa de salida cada neurona tiene una representación geométrica como un cuadrado o una celda hexagonal. Estas celdas pueden después colorearse, de acuerdo a algún código para representar varios tipos de resultados anaíiticos.

Los escenarios de visualización de la red neuronal SOM proveen una forma novedosa de visualizar el conocimiento encriptado en los datos de una manera muy original, distinguiéndose de los escenarios gráficos que clásicamente han sido utilizados por Ingenieros, Físicos y Matemáticos.

En la sección se describen formalmente los elementos esenciales que constituyen la familia de algoritmos SOM 2.1. En las secciones 2.2 *SOM Básico* y 2.2.2 *Batch Map* se definen y discuten los dos miembros de la familia SOM más usados. Por último, en 2.3 se establecen los conceptos analíticos para evaluar, cuantitativamente, el desempeño de estos algoritmos.

2.1. Esquema General

Las diferentes variantes de la familia de algoritmos SOM han surgido para satisfacer los requerimientos de los diferentes dominios de aplicación. Estas distintas variantes pueden implicar cambios, ajustes y mejoras de diferente naturaleza al método básico.

Por ejemplo, la naturaleza del universo de entrada requiere utilizar retículas con geometrías variadas en la capa neuronal de salida o definir apropiadas medidas de similitud. También se puede requerir el uso de diferentes técnicas para procesar señales, series de tiempo, cadenas de símbolos e incluso entradas heterogéneas.

Los ajustes también pueden implicar variaciones funcionales para integrar criterios de competencia alternativos y aprendizaje supervisado. Algunas aplicaciones requieren un cómputo intensivo y es necesario optimizar los algoritmos de entrenamiento para acelerar el proceso computacional.

Definiciones Generales

La forma más general de los modelos de la familia SOM queda definida por la modificacón en algunos de sus componentes, de manera que una descripción

general se puede establecer de la siguiente manera,

$$SOM(\mathbb{X}, X, \mathcal{N}, \mathcal{T}, T, W, \Delta, c, r). \tag{2.1}$$

Cada uno de los elementos que lo componen se describen a continuación:

- $\mathbb{X}$ es el *universo muestral* en el cual se han modelado matemáticamente los datos. $\mathbb{X}$ es un espacio de disimilitud en el sentido definido en la sección 1.1.4. Frecuentemente, se considera $\mathbb{X}$ como un espacio euclidiano n-dimensional

$$\mathbb{X} = \mathbb{R}^n. \tag{2.2}$$

- $X \subset \mathbb{X}$ está constituido por una colección de N datos con los cuales se va a entrenar la red neuronal y se le llama *conjunto de entrenamiento*:

$$X = \{x_1, x_2, ..., x_N\}. \tag{2.3}$$

- $\mathcal{N}$ representa la arquitectura de la red neuronal, la cual se concibe constituida por varias capas de neuronas. En las aplicaciones que consideraremos, $\mathcal{N}$ está constituida por dos capas, la capa de entrada y la capa de salida: $\mathcal{N}_I, \mathcal{N}_O$. También consideraremos que:

$$\#\mathcal{N}_I = \dim(\mathbb{X}) = \dim(\mathbb{R}^n) = n. \tag{2.4}$$

- La capa de procesamiento $\mathcal{N}_O$ se considerará constituida por un rectángulo retícular estructurada de tal manera que cada elemento de la retícula corresponde a una neurona. Las retículas más usadas son las regulares (cuadrada o hexagonal). Se considera que, la retícula modela una red de k neuronas y en un inicio a cada neurona η de la retícula se le asocia aleatoriamente un vector de referencia $w_0^\eta \in \mathbb{X}$, de manera que,

$$W_0 = \{w_0^\eta\}_{\eta \in \mathcal{N}_O} \tag{2.5}$$

$$\#W = \#\mathcal{N}_O = k. \tag{2.6}$$

  Resulta conveniente considerar que el conjunto de vectores de referencia W tiene una estructura matricial y que entonces $W \in \mathbb{R}^{k \times n}$.

- El proceso de entrenamiento de la red consiste en modificar la matriz W adaptivamente para producir la secuencia: $\{W_0, ..., W_\Omega\}$. Donde Ω representa el tiempo de paro del entrenamiento para las neuronas de $\mathcal{N}_O$. Por lo que el conjunto $\mathcal{T} = \{0, ..., \Omega\} \subset \mathbb{N}$ es entonces el conjunto de tiempos de procesamiento, donde $t = 0$ será el tiempo donde se escojen aleatoriamente los vectores de referencia iniciales y $t = 1$ es el tiempo donde se realiza la primera actualización y $t = \Omega$ el tiempo de la última actualización.

- Durante el entrenamiento los datos de entrada se presentan a la red en κ ciclos y cada ciclo de presentación del conjunto de datos de entrada es llamada una época del entrenamiento. Así pues, durante el entrenamiento la capa de entrada recibe $\lambda = N\kappa$ presentaciones de datos. Sea

$$T = \{1, ..., \lambda\} \subset \mathbb{N} \tag{2.7}$$

el conjunto de tiempos de presentación de los datos.

- La función $x : T \to X$ representa el proceso de selección de los estímulos (datos) que son presentados a las neuronas de la capa de entrada. Este proceso puede ser determinista o estocástico. Una opción muy socorrida es ordenar aleatoriamente el conjunto X y presentar los datos secuencialmente en el orden establecido y repetir este proceso durante κ épocas. Así para cada $\tau \in T$ se tiene:

$$x(\tau) = x_{(\tau \mod N)+1} \tag{2.8}$$

- Sea $d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ es una función de disimilitud definida como en (1.3). Se define la función $c : \mathbb{R}^{k \times n} \times \mathbb{X} \rightharpoonup \mathcal{N}_O$ que es el criterio de competencia. Dado $\mathbb{R}^{k \times n}$, para cada $x \in \mathbb{X}$ se determina la neurona ganadora $c(W, x) \in \mathcal{N}_O$ como:

$$d(w_{c(W,x)}, x) = \text{mín}\left\{d(w_\eta, x)\right\}_{\eta \in \mathcal{N}_O} \tag{2.9}$$

Para el caso ambiguo en el que existan al menos dos $\eta, \eta' \in \mathcal{N}_O$ y $x \in \mathbb{X}$ tales que:

$$d(w_\eta, x) = \text{mín}\left\{d(w_\eta, x)\right\}_{\eta \in \mathcal{N}_O} = d(w_{\eta'}, x),$$

$c(W, x)$ se puede resolver la ambiguedad mediante una selección aleatoria.

- Durante el entrenamiento es práctico referirse a la neurona ganadora en el tiempo $\tau \in T$ y el estado de los pesos sinápticos W_t de manera que se define:

$$c(\tau) = c(W_t, x(\tau)) \tag{2.10}$$

En el siguiente apartado se aclarará la relación entre t y τ y se define una función de acoplamiento $t(\tau)$.

- La función $\Delta : \mathcal{N}_O \times \mathcal{N}_O \times \mathcal{T} \rightharpoonup \mathbb{R}$ establece los pesos de las conexiones entre las neuronas de $\mathcal{N}_O$ en cada mometo de actualización t. Lo más usual es que esta función esta predeterminada es decreciente como

función de t y de la distancia $d_{\mathcal{N}_O}$ entre las neuronas de $\mathcal{N}_O$. Dados $\eta, \nu \in \mathcal{N}_O$ y $t \in \mathcal{N}$, queda definida como:

$$\delta^t_{\eta\nu} = \Delta(\eta, \nu, t) \tag{2.11}$$

- Se define $R : \mathbb{R}^{k\times n} \times \mathcal{T} \to \mathbb{R}^{k\times n}$ como la regla de aprendizaje. En el escenario más simple tiene la siguiente forma:

$$\begin{aligned} W_{t+1} &= R(W_t, \tau) \\ &= W_t - \Delta(W_t - [x(t)]) \end{aligned} \tag{2.12}$$

  Donde $[x(\tau)] \in \mathbb{R}^{k\times n}$ es tal que sus renglones son copias de $x(\tau)$. Si se define neurona a neurona 2.12 tiene la siguiente forma:

$$w_\eta(t) + \delta^t_{\eta,c(t)}(x(\tau) - w_\eta(t)) \tag{2.13}$$

Estrategias Computacionales

Tradicionalmente se han considerado dos estrategias computacionales para el entrenamiento, conocidos como: el SOM Básico y *Batch Map*. Como su nombre lo indica, el SOM Básico constituye el esquema más símple, pero el Batch Map es computacionalemente más eficiente.

Para crear un esquema formal, que abarca ambas estrategias, se define la función $t : T \to \mathcal{T}$ de manera que a cada tiempo de presentación $\tau \in T$ le asocia $t \in \mathcal{T}$ como el tiempo de actualización de los pesos sinápticos que corresponde a τ. Para el caso del Básico $T = \mathcal{T}$ y

$$t(\tau) = \tau. \tag{2.14}$$

El *Batch Map* es computacionalmente más eficiente porque se puede implementar en arquitecturas de hardware con múltiples procesadores. En este caso las actualizaciones de los pesos sinápticos no se realizan con cada presentación de los datos, sino se pospone la actualización hasta que son presentados todos los datos del conjunto X. Así, la función de acoplamiento queda definida de la siguiente manera.

$$t(\tau) = \lfloor \frac{\tau}{N} \rfloor, \tag{2.15}$$

donde los paréntesis $\lfloor . \rfloor$ representan la función menor entero.

2.2. Algoritmos de Entrenamiento

El SOM Básico es el punto de partida de los modelos de la familia SOM Kohonen [1982b]. En este modelo se considera una arquitectura $\mathcal{N}$, con una capa de entrada básica $\mathcal{N}_I$ y una capa de procesamiento $\mathcal{N}_O \subseteq \mathbb{R}^2$, la cual puede estar organizada en una malla hexagonal o cuadrada (la Figura 11 representa la arquitectura del SOM básico. La mayoría de los modelos de la Familia SOM identificados en la literatura tienen esta arquitectura. A continuación se detalla el algoritmo de entrenamiento.

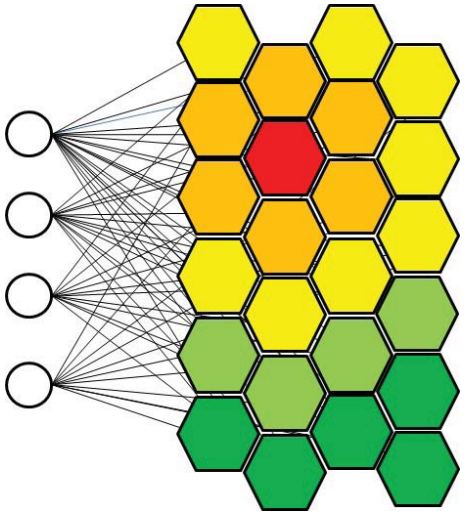


Figura 11: Arquitectura del SOM Básico

2.2.1. El SOM Básico

El proceso de entrenamiento SOM Básico propuesto originalmente, considera los casos más simples para: el universo muestral (2.2) y la presentación de los datos es periódica (2.8). Para el criterio de competencia (2.9) utiliza como índice de disimilitud la métrica euclidiana.

$$d(x, w_\eta) = \sqrt{\sum_{i=1}^{n} (x^i - w_\eta^i)^2}. \tag{2.16}$$

La forma de la regla de aprendizaje es como en (2.12) y se aplica con tiempo de iteración de la forma (2.14) y tiempo de terminación Ω.

Data: $X \subset \mathbb{R}^n$, $W_0 \in \mathbb{R}^{k \times n}$ aleatorio
Result: $W_\Omega \in \mathbb{R}^{k \times n}$
for $t \leftarrow 1$ **to** Ω **do**
  $c(t) = c(W_{t-1}, x(t))$
  forall $\eta \in \mathcal{N}_O$ **do**
    $w_\eta(t) = w_i(t-1) + \delta^t_{c(t)\eta}(x(t) - w_i(t-1))$
  end
end

Algorithm 1: SOM Básico

Función Δ

El factor de ajuste Δ tiene la forma:

$$\delta^t_{c(t)\eta} = \alpha(t) H_{\rho(i)}(d_{\mathcal{N}}(\eta_{c(t)}, \eta_i)), \tag{2.17}$$

donde las funciones: α, H y ρ cumplen las condiciones que se especifican en los párrafos siguientes.

Factor de aprendizaje α

La función $\alpha(t)$ que aparece en (2.17) se conoce como *factor de aprendizaje* y se le deben requerir las siguientes condiciones para garantizar la convergencia:

- $0 < \alpha(t) < 1$,
- $\alpha(t)$ no creciente,
- $\alpha(t) \longrightarrow 0$ cuando $t \longrightarrow \infty$.

una opción sugerida por KohonenKohonen [1995] es la siguiente:

$$\alpha(t) = \begin{cases} 0{,}9(1 - \frac{t}{\Omega}) & t < \Omega \\ 0 & t \geq \Omega \end{cases}. \tag{2.18}$$

Función Vecindad H

La función $H_{\rho(t)}(d_{\mathcal{N}}(\eta_{c(t)}, \eta_i))$, denominada *función vecindad*, toma valores en $[0, 1]$. Independientemente de cual sea su forma explícita, debe ser tal que $H_{\rho(t)}(0) = 1$ para todo t (inicialmente el factor de ajuste para la neurona

ganadora debe ser cercano a 1); además, para cada t fijo, $H_{\rho(t)}(d_{\mathcal{N}}(\eta_{c(t)}, \eta_i))$ debe ser decreciente como función de $d_{\mathcal{N}}(\eta_{c(t)}, \eta_i) \in \mathbb{R}^+$ y cumplir con

$$H_{\rho(t)}(d_{\mathcal{N}}(\eta_{c(t)}, \eta_i)) \to 0, \tag{2.19}$$

cuando $d_{\mathcal{N}}(\eta_{c(t)}, \eta_i)$ se incrementa. Una forma común para H es la siguiente:

$$H_{\rho(t)}(d_{\mathcal{N}}(\eta_{c(t)}, \eta_i)) = \exp(-\frac{d_{\mathcal{N}}(\eta_{c(t)}, \eta_i)^2}{2\rho^2(t)}), \tag{2.20}$$

donde $\rho(t)$ corresponde al ancho promedio de la vecindad alrededor de la neurona ganadora $\eta_{c(t)}$. Por último, para efectos de la convergencia del algoritmo, $\rho(t)$, debe cumplir:

- $t \leq t' \Longrightarrow \rho(t) \geq \rho(t')$,
- $\rho(t) \to 0$ cuando $t \to \infty$.

Se recomienda que $\rho(0) = \rho_{max}$ se escoja más grande que la mitad del diámetro de la red. Kohonen propone lo siguiente Kohonen [1995]:

$$\rho(t) = \begin{cases} \rho_{\text{máx}}(1 - \frac{t}{\Omega}) & t < \Omega \\ 0 & t \geq \Omega \end{cases} \tag{2.21}$$

2.2.2. El Batch Map

Una variante del algoritmo SOM es el algoritmo *Batch Map* (BMSOM). Éste modelo comparte características con el SOM Básico, como son: arquitectura $\mathcal{N}$ de la red neuronal, propiedades (2.2), (2.8) y tiempo de terminación Ω.

En el algoritmo SOM Básico, cada estímulo provoca una actualización, mientras que, en el BMSOM se presentan todos los datos y posteriormente se realiza la actualización de los vectores de referencia. El SOM Básico realiza su actualización considerando únicamete la información provista por un dato, en cambio, el BMSOM considera el resultado de la presentación de todos que corresponden a una época.

Este cambio abre la posibilidad de implementar el entrenamiento de la red neuronal SOM en esquemas paralelizables. En consecuencia, hay una reducción significativa del tiempo de entrenamiento, pero no se observan cambios significativos en los resultados finales obtenidos mediante el BMSOM y el SOM Básico Cheng [1997].

La principal diferencia entre estos métodos de entrenamiento está en la regla de actualización y la sincronización del tiempo de presentación de los

datos de entrada con el tiempo de la capa de actualización. Como hemos visto antes, el BMSOM considera una función de sincronización de la forma $t : T \to \mathcal{T}$ de la forma:

$$t(\tau) = \lfloor \frac{\tau}{N} \rfloor.$$

Entonces el algoritmo de entrenamiento toma la forma que se muestra en Algoritmo 2.

Data: $X \subset \mathbb{R}^n$, $W_0 \in \mathbb{R}^{k \times n}$ aleatorio
Result: $W_\Omega \in \mathbb{R}^{k \times n}$
forall $\eta \in \mathcal{N}_O$ **do**
  $V_\eta = \emptyset, \mu_\eta = w_\eta(0)$
end
for $\tau \to 1$ **to** λ **do**
  $c(\tau) = c(W_{t(\tau)}, x(\tau))$
  $V_{c(\tau)} = V_{c(\tau)} \cup \{x(\tau)\}$
  if $t(\tau + 1) > t(\tau)$ **then**
    forall $\eta \in \mathcal{N}_O$ **do**
      if $V_\eta \neq \emptyset$ **then**
        $\mu_\eta = \frac{1}{\#V_\eta} \sum_{x \in V_\eta} x$
      end
    end
    forall $\eta \in \mathcal{N}_O$ **do**
      $w_\eta(t(\tau)) = \frac{\sum_{\nu \in \mathcal{N}_O} \#V_\nu H_{t(\tau)}(\nu, \eta, t(\tau))\mu_\nu}{\sum_{\nu \in \mathcal{N}_O} \#V_\nu H_{t(\tau)}(\nu, \eta)}$
    end
    forall $\eta \in \mathcal{N}_O$ **do**
      $\#V_\eta = 0, \mu_\eta = w_\eta(t(\tau))$
    end
  end
end

Algorithm 2: Batch Map

2.3. Preservación Topológica

La proyección de los datos al arreglo bidimensional de neuronas, queda definido de la siguiente manera:

$$\varphi : X \to \mathcal{N}_O, \tag{2.22}$$

$$\varphi(x) = c(W_\Omega, x), \tag{2.23}$$

donde $c(x) = c(W_\Omega, x) \in \mathcal{N}_O$ es la neurona ganadora del dato $x \in \mathbb{X}$.

En general, la manera en que los datos se distribuyen en la malla queda determinada por las regiones del espacio en el cual se especializan las neuronas

$$V_\eta = \{x \in X \mid \varphi(x) = \eta\}, \tag{2.24}$$

este conjunto es denominado **conjunto de Voronoi**.

Nótese que $X = \bigcup_{\eta \in \mathcal{N}_O} V_\eta$, ya que para $x \in X$ la función $d(x, \quad)_{|W}$ tiene, al menos un w mínimo en W (vector de referencia más cercano) por lo tanto $x \in V_\eta$. Procediendo de manera análoga, si se consideran todos los elementos del universo, cada $w_\eta \in W$ determina una región de $V_\eta \subseteq \mathbb{X}$ denominada **región de Voronoi**:

$$\hat{V}_\eta = \{x \in \mathbb{X} \mid \varphi(x) = \eta\}, \tag{2.25}$$

Nótese que $\hat{V}_\eta$ contiene a V_η y que $\mathbb{X} = \bigcup_{v \in W} \hat{V}_v$.

La proyección que efectúa el mapeo ϕ, sobre la capa de salida de la red neuronal representará las relaciones de similitud de los datos de entrada mediante una relación de cercanía en $\mathcal{N}_O$, así:

$$\text{si } s(x, y) \text{ es grande entonces } d(\phi(x), \phi(y)) \text{ es pequeño.} \tag{2.26}$$

A la propiedad (2.26) se le suele denominar **preservación de la topología**. Se considera esta propiedad como la más importante del SOM, ya que permite utilizar a los mapas como representaciones visuales de las relaciones de similitud entre los datos.

Para cuantificar la calidad del mapeo inducido ϕ, se han utilizado distintas medidas. La más empleada es la propuesta de Villmann y colaboradores Villmann et al. [1994]. La medida considera la adyacencia entre las neuronas y los conjuntos de Voronoi; y que tanto se preservan esas cercanías en la red. La función del error topográfico se define de la siguiente manera:

$$\Psi(r) = \frac{1}{k} \sum_{\eta \in \mathcal{N}_O} \#\{\eta \in \mathcal{N}_O | d_{\mathcal{N}_O}(\hat{\eta}, \eta) \geq r \wedge \hat{V}_{\hat{\eta}} \cap \hat{V}_\eta \neq \emptyset\} \tag{2.27}$$

Observación 1 *Ψ es monótona decreciente, por lo que, si existe r_0 tal que $\Psi(r_0) = 0$ entonces $\forall r \geq r_0$, $\Psi(r) = 0$.*

Cuando $\Psi \equiv 0$ se dice que, la **preservación de topología** es perfecta. En las aplicaciones rara vez se tiene la preservación perfecta. Es útil considerar como indicador de bondad el que Ψ se aproxime rápidamente a 0 cuando r crece.

2.4. Etapas del entrenamiento

Este proceso tiene dos etapas: *ordenamiento global y refinamiento.*

- **Ordenamiento Global:** En esta etapa se ajustan los pesos de cada una de las neuronas, para localizar las regiones del espacio de entrada que están más pobladas por el conjunto de datos X. Durante esta fase no es deseable que los valores de $\alpha(t)$ y $\rho(t)$ sean muy pequeños.
- **Refinamiento:** Después de la fase de ordenamiento, conviene que los valores de $\alpha(t)$ y $\rho(t)$ se vayan haciendo pequeños para alcanzar una mayor precisión. Esta es la etapa más larga del proceso porque requiere varias épocas de entrenamiento.

Para ilustrar el proceso de entrenamiento e introducir elementos analíticos que posteriormente serán de utilidad, mostramos un experimento donde, la red SOM identifica la estructura de un conjunto de datos en el espacio euclideano tridimensional. Estos datos están organizados a priori en un conjunto de 10 bolas de diferente tamaño (ver figura 12). El experimento se realiza con las siguientes especificaciones:

- Un conjunto de entrenamientos $X \subset \mathbb{X}^3$ que consta de 10 conjuntos de mil puntos cada uno, distribuidos uniformemente en bolas que no se intersectan entre sí.
- modelo SOM Básico con capa de salida $\mathcal{N}_O$ de neuronas desplegadas en una retícula bidimensional hexagonal.
- W_0 se determina uniformemente distribuido en el cubo unitario.
- $\Omega = 50,000$.
- Δ es tal que α tiene la forma (2.18), H tiene la forma (2.20) y ρ tiene la forma (2.21).

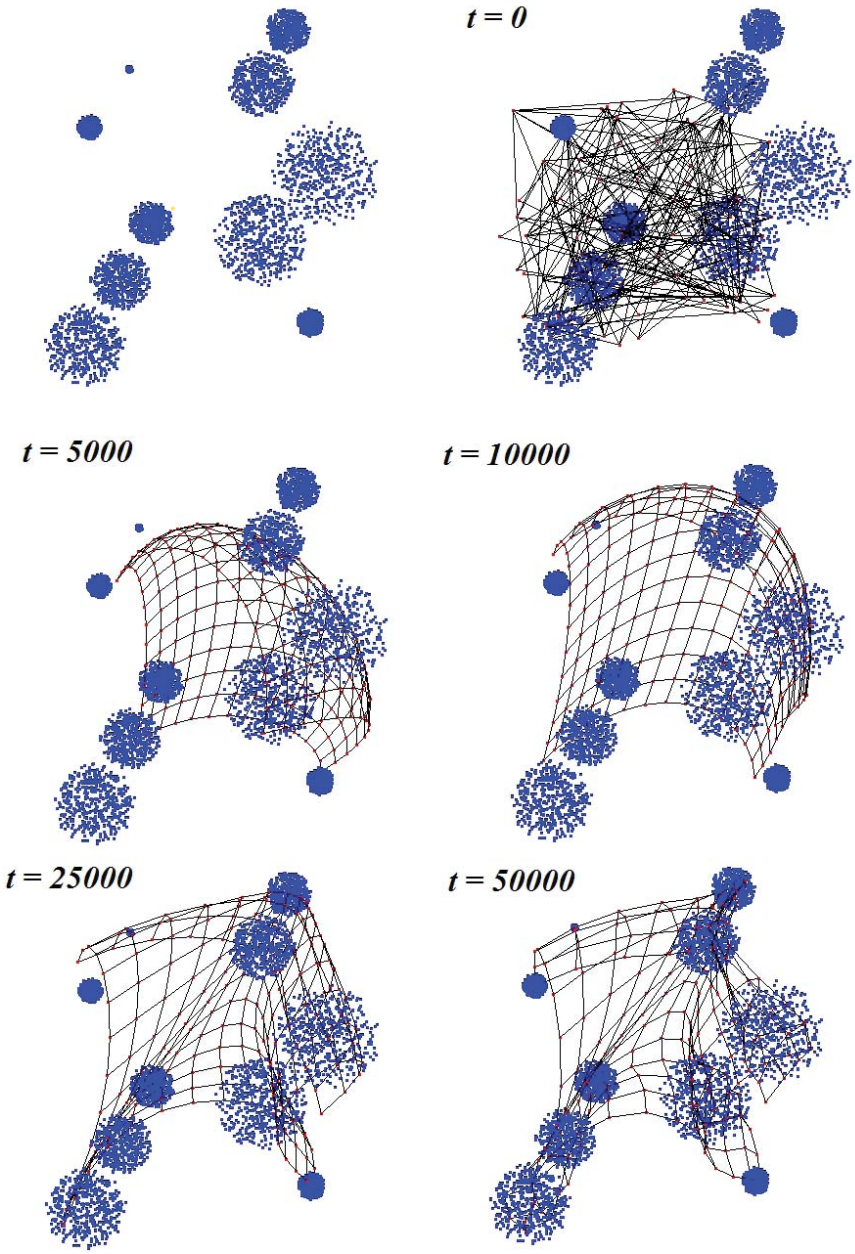


Figura 12: Evolución del entrenamiento de una red SOM.

La figura 12 ilustra varias fases del proceso de entrenamiento. En la primera parte de la gráfica[1] se muestra como el conjunto X distribuido en las bolas, en la segunda parte ($t = 0$) se agregan los vectores de referencia $\{w_\eta(0)\}_{\eta\in\mathcal{N}_O} \in \mathbb{R}^3$. Estos vectores se grafican como pequeños puntos rojos los cuales están unidos entre sí. La adyaciencia de la red que forma el conjunto de $\{w_\eta(0)\}_{\eta\in\mathcal{N}_O}$ está dada por la adyacencia entre las neuronas en la retícula. La condición inicial aleatoria de los vectores de referencia se hace muy notoria cuando se observa la "maraña"que se forma en el la gráfica $t = 0$. Las gráficas sucesivas son distintos momentos en el entrenamiento $t = 5000, 10000, 25000$ hasta llegar a $t = \Omega$.

La etapa de ordenamiento global se observa en los momentos $t = 5000, 10000$. Un indicador visual de esto es cuando la malla se ve bastante regular en estas primeras etapas. Conforme avanza el entrenamiento, la malla se va deformando para adaptarse a estructuras cada vez más locales. Estas deformaciones, se observan en $t = 25000$. La última imagen muestra como muchos de los vectores de referencia se han adaptado para quedar distribuidos en los clusters de puntos definidos por las esferas. Unos cuantos vectores de referencia quedan destinados a representar el espacio vacío existente entre los clusters.

Las imágenes de la figura 12 fueron creadas usando el sistema de software LabSOM Jiménez-Andrade et al. [2007], Carrillo et al. [2011], desarrollado en colaboración con el grupo de trabajo del *Laboratorio de Dinámica no-Lineal* (LDNL), de la *Facultad de Ciencias* de la *UNAM* y constituye un producto tecnológico asociado al proyecto de investigación que en esta tesis se reporta.

[1] Se considera el orden de arriba hacia abajo, de izquierda a derecha

Capítulo 3

Minería de Datos Informétrica con la Red Neuronal SOM

Los métodos basados en algoritmos SOM resultan útiles en tareas típicas de minería de datos como:

- Cuantización Vectorial Somervuo and Kohonen [1999].
- *Clustering* Vesanto and Alhoniemi [2000].
- Proyección no-lineal Venna and Kaski [2001], Wu and Chow [2005], Xu et al. [2011].
- Análisis de la dinámica de sistemas complejos Principe et al. [1998], Ferreira and Araújo [2016].
- Procesamiento de datos no-vectoriales (incluyendo variables categóricas) Nikkilä et al. [2002], Olteanu and Villa-Vialaneix [2016].
- Visualización de Información Skupin [2000], Skupin and Fabrikant [2003].

Existen aplicaciones exitosas de la red neuronal en el campo de los estudios cienciométricos. Por ejemplo, las primeras investigaciones que se reportan en la literatura se utiliza el análisis de matrices de co-ocurrencias en dominios biomédicos Sotolongo et al. [2002] y en el mapeo de la ciencia y tecnología Polanco et al. [2001]Bote et al. [2002]. En este último trabajo, se utilizan arquitecturas Multi-SOM con varias capas intermedias de procesamiento. Aplicaciones más recientes, incluyen el análisis de la evolución de dominios de conocimiento Guzmán et al. [2010] y las mejoras significativas en el terreno del mapeo de la ciencia Skupin et al. [2013].

La red neuronal SOM, es muy apreciada por sus novedosas capacidades de visualización, en estas aplicaciones se aprovechan las capacidades del SOM

para generar visualizaciones que facilitan la representación y descubrimiento de nuevo conocimiento. A continuación, se revisan algunos de los métodos de visualización basados en el SOM.

3.1. Análisis visual de datos basado en el mapeo auto-organizante

En esta sección se describen varias técnicas de visualización basados en modelos SOM. El punto de partida para la aplicación de estas técnicas es el mapeo $\varphi : X \to \mathcal{N}_O$, el cual asocia vectores multidimensionales a neuronas en la malla neuronal y queda determinado una vez que se ha completado el entrenamiento.

Los métodos de visualización basados en el SOM, se pueden dividir en tres categoría Vesanto [1999]:

- Los que búscan obtener una idea global de la estructura del conjunto de datos.
- Los que se basan en el análisis de los vectores de referencia en la búsqueda de una descripción de los clusters o identificar correlaciones entre las variables.
- Los que examinan muestras de datos para clasificarlos o identificar novedades.

Las técnicas de visualización presentadas en esta tesis quedan determinadas por una función. Esta función asigna color a los hexágonos que representan a las neuronas en la malla. Se parte de un conjunto de colores C y se establece una función

$$col : \mathcal{N}_O \to C. \tag{3.1}$$

Posteriormente, cada hexágono (neurona) sobre la malla η se colorea de $col(\eta)$.

En las siguientes subsecciones, se muestran dos ejemplos clásicos de métodos de visualización basados en el SOM que corresponden a la primera y segunda categorías descritas: *U-Matrix* y *Mapas de Componentes*. En el capítulo 4 se muestra una aplicación de la tercera categoría.

3.1.1. Mapas de Componentes

Esta técnica es útil para analizar la relación existente entre las variables que representan a los datos en el espacio multidimensional $\mathbb{X}$. Facilita la

identificación de conjuntos de variables correlacionadas entre sí. Esto resulta relativamente fácil porque los descubrimientos se realizan mediante la inspección visual de mapas asociados a cada una de las variables. Las variables correlacionadas se caracterizan por compartir una distribución similar de valores. Las componentes de los vectores representan a los datos en el espacio multidimensional.

En general, esta técnica considera que el espacio de datos $\mathbb{X}$ es algún subconjunto de $\mathbb{R}^n$. Para cada una de las n componentes $\{X_i\}_{i=1:n}^n$ de los vectores en el espacio de los datos $\mathbb{X} \subseteq X_1 \times ... \times X_n \subseteq \mathbb{R}^n$, se generará una cartografía que es llamado "el mapa del i-ésimo componente".

Estos mapas de componentes se colorean definiendo una función col_i entre el intervalo de variación de cada componente x_i de $x \in X$ mapeado sobre una barra cromática $\mathcal{C}$ ordenada de acuerdo a la longitud de onda de cada color

$$col_i : [\min_{x \in X}\{x_i\}, \max_{x \in X}\{x_i\}] \to \mathcal{C}. \tag{3.2}$$

Así por cada componente x_i de los vectores en $\mathbb{X}$ se produce un mapa coloreado (Mapa de Componente) que indica cómo se distribuyen sobre la malla neuronal los valores de esta variable.

3.1.2. U-Matrix

Otra alternativa de visualización es la denominada *U-Matrix*. Las cartografías resultantes de la aplicación de esta técnica resultan de utilidad para identificar patrones de agrupamientos, usando medidas locales de conectividad entre las neuronas y la información que proporcionan los vectores de referencia.

La aplicación de la técnica parte de la distancia promedio entre el vector de referencia de una neurona y los vectores de referencia de cada una de las neuronas adyacentes sobre la malla neuronal. Los valores obtenidos para cada neurona se resaltan en el mapa usando algún patrón de coloración.

Para calcular la distancia promedio correspondiente a cada $\eta \in \mathcal{N}_O$ se considera U_η una vecindad de η y la distancia promedio u_η entre el vector de referencia ω_η y los vectores de referencia de las neuronas en U_η,

$$u_\eta = \frac{1}{\#U_\eta} \sum_{\nu \in U_\eta} |\omega_\eta - \omega_\nu| , \tag{3.3}$$

donde $\#U_\eta$ representa la cardinalidad de la vecindad.

La función de coloración *col* biyecta una barra bicromática (frecuentemente del amarillo al rojo) con el intervalo

$$[\min\{u_\eta\}_{\eta \in \mathcal{N}_O}, \max\{u_\eta\}_{\eta \in \mathcal{N}_O}].$$

El método U-Matrix, utiliza a u_η como una medida de la similitud de los datos que fueron proyectados alrededor η. El patrón de coloración originado revela el grado de cercanía, que tienen en el espacio multidimensional los conjuntos de datos asignados a neuronas muy inmediatas en el mapa.

3.1.3. Mapa de Conglomerados (Clustering)

Concluido el proceso de entrenamiento de una red SOM, es deseable representar sobre la red neuronal, el patrón de agrupamiento que tiene el conjunto de datos $X \subset \mathbb{X}$ en el espacio multidimensional. Para lograr obtener resultados, se aplica un método estándar de clustering que permita la determinación de una partición constituida por conglomerados de vectores de referencia en el conjunto $W \subset \mathbb{X}$. De manera natural esta partición de vectores de referencia $W = \{w_\eta\}_{\eta \in \mathcal{N}_O}$ determina una partición del conjunto de neuronas que pertenecen a la capa de neuronas $\mathcal{N}_O$ de la red SOM:

$$\mathcal{C} : \mathcal{N}_O \to \{1, ..., K\} \tag{3.4}$$

Esta partición de la red neuronal $\mathcal{C} = \{\mathcal{N}_j\}_{j=1}^{K}$, constituye lo que se conoce en la literatura como: clustering del mapa de neuronas. Cada conjunto $\mathcal{N}_j$ constituye un *cluster* de neuronas y para construir un *mapa de clusters* se selecciona un grupo de colores: $\zeta = \{\zeta_1, ..., \zeta_K\}$ y se usa una función de selección de color de la siguiente manera:

$$col(\eta) = \zeta_{\mathcal{C}(\eta)} \tag{3.5}$$

3.2. Métodos SOM para grandes colecciones de documentos

Métodos SOM basados en VSM

En la literatura se observan algunas aplicaciones exitosas del agrupamiento y la visualización de colecciones de documentos con redes SOM. En la mayoría de estas aplicaciones se utiliza como modelo de representación el VSM (Vector Space Model, ver sección 1.1.2).

El modelo WebSOM destaca por el volumen de documentos que procesa Kohonen et al. [2000]. Este método fue desarrollado con el objetivo de visualizar grandes conjuntos de páginas Web. El método incorpora varias técnicas para solucionar el problema de la alta dimensión, inherente al modelo de representación vectorial de los documentos. Estas técnicas incluyen: el pesado

de palabras basado en entropía, un modelo SOM con una arquitectura de multi-capa y un mapeo disperso aleatorio (RSM).

Otro ejemplo de la aplicación de modelos SOM empleando VSM se presenta en un estudio de Pullwitt Pullwitt [2002]. El método incorpora técnicas de análisis semántico en el proceso de representación vectorial de los documentos. Para ello, utiliza un método de clasificación de oraciones, la cual considera información contextual y semántica.

Correa y Ludemir, proponen un método de representación de documentos que, incorpora relaciones semánticas en la representación vectorial de un conjunto de *términos*, estos vectores se utilizan para entrenar una red SOM Corrêa and Ludemir [2006]. Como resultado del entrenamiento se forman grupos de términos co-ocurrentes. La proyección inducida de palabras clave se denomina *Mapa Semántico* (SM). Los autores construyen una matriz basados en el Mapa Semántico para realizar una proyección lineal similar a la que define el RMS. Se demuestra que, la metodología basada en el Mapa Semántico ofrece mejores resultados que los obtenidos por el RMS.

Se presenta una mejora del método SM gracias a un paso de reducción de volumen. Este consiste en el uso de un algoritmo de agrupamiento de las representaciones vectoriales de los documentos Corrêa and Ludermir [2008]. Estos desarrollos han sido aplicados en el área de procesamiento de lenguaje natural.

Un análisis de los trabajos mencionados permiten derivar las siguientes conclusiones::

- Existen limitantes intrínsecas en el plantemaiento del SOM original que lo hacen difícil de aplicar en casos donde la dimensión crece en función de la cantidad de datos ($n \uparrow$ cuando $N \uparrow$).

- Se obtienen mejores resultados en la medida en que se incorporan elementos que agregan significado (semánticos) a la represcentación de documentos.

En el Capítulo 6 se propone un método que incorpora el etiquetado semántico de documentos (descriptores ordenados en una ontología) y los patrones temporales de las ocurrencias de los descriptores para realizar el SM y analizar el panel informétrico $\{D_t\}_{t\in\mathbb{T}}$. Esta representación permite interpretar la proyección de documentos. Se consideran los aspectos dinámicos de la información expresada en la ocurrencia temporal de los descriptores. Específicamente, el mapeo de los documentos se realiza considerando un modelo de gráfica bipartita en lugar del modelo vectorial.

Métodos SOM en modelos de gráficas

Durante la década pasada se propusieron métodos para analizar colecciones de documentos basados en el SOM que, incorporan representaciones basadas en gráficas.

Phuc y Hung, en su propuesta de modelo SOM consideran como dato de entrada un conjunto de gráficas de palabras clave Phuc and Hung [2008]. Esta variante, incorpora una función de disimilitud para las gráficas basadas en el concepto de subgráfica común máxima y una regla de aprendizaje que utiliza un algoritmo genético para buscar las gráficas comunes. Este enfoque es conceptualmente interesante, pero con una alta complejidad computacional, lo cual lo hace inviable para su aplicación en el procesamiento de grandes colecciones de documentos.

Una forma de simplificar la representación es considerar árboles. Se ha observado que, para procesar datos estructurados en forma de árbol de una manera eficiente, es equipando la arquitectura del SOM con conexiones de retroalimentación como lo han planteado varios autores Chappell and Taylor [1993], Koskela et al. [1998], Voegtlin [2002], Strickert and Hammer [2005] y Hagenbuchner et al. [2003]. Un ejemplo de estos modelos es el SOM-SD (*SOM for structured data*) .

El SOM-SD, propuesta de Hagenbuchner y otros autores, produce mejores resultados al comparar con los modelos mencionados, en términos de una métrica de profundidad del árbol Hagenbuchner et al. [2003] Vanco and Farkas [2010]. Este modelo ha sido empleado con éxito para el agrupamiento de documentos XML. Ventaja que le permitió ganar la competencia internacional *Iniciativa para la evaluación de la recuperación de XML* (INEX) en el año 2006 y 2007.

En el capítulo 6 se propone un mapeo de documentos basado en una representación de gráfica bipartita y en el concepto de Mapa Semántico, que además aprovecha la informacioń obtenida de la ocurrencia temporal de los descriptores.

3.3. Metodología ViBlioSOM

Aunque la principal aplicación del SOM en el contexto del análisis informétrico es para realizar *Science Mapping*. Sin embargo, poco se ha hecho en la aplicación de esta herramienta al análisis de otros datos cienciométrico como indicadores de productividad e impacto (eficacia y eficiencia), en el desempeño de unidades de producción a distintos niveles de agregación (micro (autores), mezo (instituciones), macro (país)).

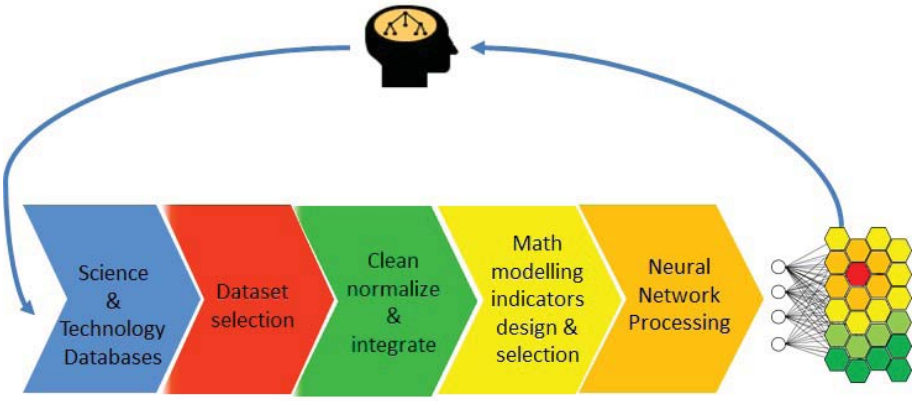


Figura 13: La metodología ViBlioSOM está orignalmente pensada para operar con indicadores bibliométricos a distintos niveles jerárquicos y en distintos espacios abstractos.

Originalmente la metodología ViBlioSOM fue diseñada para el descubrimiento de conocimiento en bases de datos de textos científicos, como parte de un proyecto de colaboración entre científicos de la información, computólogos y matemáticos. Originalmente, se propuso el uso secuencial de sistemas de software propietarios Sotolongo et al. [2002] para llevar a cabo el proceso KDD Fayyad et al. [1996]. La idea es calcular conjuntos de representaciones vectoriales basadas en indicadores bibliométricos, que son analizados en la fase de minería de datos usando redes neuronales SOM.

La aplicación secuencial de sistemas de software para el procesamiento, análisis y visualización de la información, resultó útil para realizar estudios cienciométricos, pero la interrupción del proceso para cambiar de un sistema de software era un inconveniente. Esto motivó, el diseño y desarrollo de un sistema de software con la funcionalidad suficiente para realizar cada una de las fases del proceso ViBlioSOM de una manera integral.

La metodología ViBlioSOM (Visualización Bibliométrica usando la red neuronal SOM)Guzmán [2010], ha sido implementada en un sistemas de software desarrollados por investigadores y tecnólogos asociados al Laboratorio de Dinámica No Lineal de la UNAM (Universidad Nacional Autónoma de México) en colaboración con científicos del Instituto Finlay.

La metodología ViBlioSOM está relacionada con las etapas generales del KDD, las cuales son:

1. Ubicación del objetivo.

2. Adquisición y selección de datos.

3. Pre-procesamiento.

4. Minería de Datos y Textos.

5. Visualización e interpretación de resultados.

6. Informe y distribución de los resultados.

Las aportaciones metodológicas de este trabajo se enmarcan en el contexto de la implementación de la metodología ViblioSOM Guzmán [2010] y representan mejoras en el planteamiento original, donde se plantean aplicaciones directas en el caso $X \subset \mathbb{R}^n$, desde la modelación de los datos como objetos abstractos ($x \in \mathbb{R}^n \times \mathbb{S}^m \times \mathbb{Z}_2^k \times \mathbb{T}$).

Parte II

Aportaciones Metodológicas

La investigación tiene como objetivo desarrollar nuevos métodos y técnicas para realizar de manera eficaz el análisis y visualización informétrica, utilizando redes neuronales de la familia SOM. La eficacia, utilidad y eficiencia de los métodos desarrollados se comprueban mediante:

- la comparación de los resultados obtenidos con los métodos propuestos contra los obtenidos utilizando versiones standar del SOM (SOM Básico Batch Map).
- la valoración del conocimiento develado por parte de especialistas del dominio de aplicación.
- la evaluación de los mapas obtenidos desde la perspectiva del experto del campo de aplicación.

La realización de este trabajo implicó identificar y abordar distintos problemas arquetípicos dentro de varios dominios de aplicación, lo cual requiere la colaboración con expertos del dominio: médicos, biólogos, científicos de la información, cienciómetras y demógrafos. Estos problemas tienen la característica que plantean dificultades generales que se presentan en diversos escenarios.

Desde la perspectiva de las ciencias de la computación, el trabajo involucró evaluar distintas soluciones (partiendo de la aplicación directa del SOM) y proponer alternativas que implican una mejora en la claridad y calidad de las visualizaciones obtenidas.

Una hipótesis muy importante, que se corrobora con los experimentos desarrollados en esta tesis doctoral, es: *las redes neuronales SOM constituyen un recurso útil para el análisis y la generación de representaciones visuales de la estructura y las relaciones que guardan los datos investigados.*

Como se mostrará en los siguientes capítulos, la propuesta standar del SOM (el SOM Básico o *Batch Map*) no siempre arroja buenos resultados. La puesta a punto de los algoritmos de entrenamiento para aplicaciones específicas es un arte que requiere una adecuada modelación de los datos y un conocimiento profundo del algoritmo SOM, para poder diseñar las métricas apropiadas y otros aspectos técnicos como los que se refieren a continuación:

- La alta dimensión de los datos puede ser una limitante dada la complejidad computacional (es cuadrática respecto a la dimensión). Por lo tanto, apremia desarrollar algoritmos eficientes computacionalmente y aprovechar esquemas de procesamiento en paralelo y arquitecturas de cómputo de alto desempeño.

- La calidad de los mapas obtenidos puede ser mejorada mediante la transformación matemática de los datos como pueden ser escalamientos o normalizaciones ad-hoc a cada clase de estudio. La aplicación de transformaciones puede revelar las relaciones espaciales de los vectores que modelan los datos y permitir una interpretación más intuitiva de los mapas.

- La complejidad estructural de los datos impone retos en la definición de los operadores del algoritmo. En particular, la definición de una medida de similitud adecuada. Se requiere, por ejemplo, diseñar métricas adecuadas para trabajar con datos híbridos (vectores con variables categóricas y variables numéricas) y/o con patrones temporales asociados.

- El desarrollo de herramientas computacionales accesibles para los analistas de información, las cuales faciliten el uso de métodos basados en la red neuronal SOM, en aplicaciones de análisis de datos del mundo real y que aprovechen los recursos computacionales disponibles.

En los siguientes capítulos se presentan varios ejemplos paradigmáticos de aplicación de la red neuronal SOM. Se escogieron diversos escenarios que demandan el desarrollo de diferentes recursos técnicos. Los escenarios seleccionados son:

- la evaluación cienciométrica de la producción científica de instituciones de educación superior.

- la identificación de diferencias impresas en el egreso universitario de alumnos de licenciatura con un enfoque de estudio de género.

- el mapeo de la evolución de dominios de conocimiento.

En todos los escenarios seleccionados, los datos primarios son conjuntos de documentos semi-estructurados, como los referidos en 1.1, que se pueden representar como un conjunto de objetos abstractos $d \in D \subset \mathbb{X}$. El tipo de dato más general que se considera en este trabajo pertenece a un conjunto, $\mathbb{X}$, que tiene esta esta forma general:

$$\mathbb{X} = \mathbb{R}^n \times \mathbb{S}^m \times \mathbb{Z}_2^k \times \mathbb{T}.$$

Donde, $\mathbb{R}^n$ es un vector numérico que puede representar mediciones o indicadores estadísticos asociados a cada registro, $\mathbb{S}^m$ es la esfera unitaria de dimensión m y se utiliza para representar espacios semánticos, $\mathbb{Z}_2^k$ en un vector de variables categóricas binarias y $\mathbb{T}_0$ es una variable temporal.

Usando esta terminología podemos describir genéricamente los tres tipos de datos que, se considerarán en los estudios de caso en la segunda parte del reporte de mi investigación:

Caracterización multiparamétrica no supervisada

$$n > 0, m = 0, k = 0, \tau = 0 \tag{3.6}$$

Análisis y visualización de datos híbridos

$$n = 1, m = 0, k > 0, \tau > 0 \tag{3.7}$$

Visualización de la evolución de dominios de conocimiento

$$n = 0, k > 0, m > 0, \tau > 0 \tag{3.8}$$

Para cada estudio de caso, se dedica un capítulo en el cual se aplica la metodología desarrollada. Los tres capítulos están estructurados de la misma manera. A continuación se describe esta estructura.

En la introducción se plantea la problemática en general y se mencionan los elementos de valor de las aportaciones. En la primera sección se presenta el caso de estudio estableciendo los datos, la fuente y el modelo de representación, al final de este capítulo, se exponen los resultados que se obtienen de la aplicación directa del SOM. En la segunda sección se presenta la variante del algoritmo SOM y las técnicas de visualización propuestas, las cuales representan las aportaciones de esta tesis. En la tercera sección se presentan los resultados obtenidos por los métodos y técnicas propuestas y se interpretan los resultados. Finalmente, se discuten estos resultados desde la perspectiva del dominio de aplicación.

Capítulo 4

Caracterización multifactorial de patrones de desempeño cienciométricos

Normalmente los criterios para medir el desempeño de entidades (e.g. industrias, empresas, instituciones, individuos, etc) se establecen a partir de parámetros numéricos (indicadores). Estos parámetos hacen referencia a distintas mediciones de la actividad productiva y a la calidad de los resultados o productos que se obtienen. El contar con múltiples parámetros asociados a una entidad productiva implica que su perfil de desempeño queda representado como un punto en un espacio multidimensional $x \in \mathbb{R}^n$.

En este capítulo se reportan aportaciones metodológicas para evaluar el desempeño de entidades productivas, utilizando múltiples criterios. El problema es comparar el desempeño de distintas unidades. Este problema es muy general porque se puede presentar en cualquier escenario donde existan distintas formas de medir el desempeño de una actividad productiva específica.

Dos casos de estudio fueron seleccionados para mostrar las ventajas de nuestra propuesta metodológica: el análisis del desempeño cienciométrico de las instituciones de educación superior mexicanas (IES) Arencibia et al. [2016] y el estudio de los perfiles de desempeño cienciométrico de las revistas mexicanas Villaseñor et al. [2017]. En las dos aplicaciones se aprovecharon los recursos biliométricos provistos por la base de datos Scopus y por el laboratorio cienciométrico de Scimago. En particular, se utiliza el cómputo de indicadores provistos por el SIR (*Scimago Institutional Rank*) y el SJR (Scimago Journal & Country Rank). Además, en el caso del análisis de la producción científica de las IES mexicanas se utilizó también el padrón de investigadores miembros del SNI, que actualiza cada año el CONACyT. Al

considerar el número de miembros del SNI de cada institución es posible computar indicadores de productividad y comparar instituciones con distintas capacidades productivas.

Específicamente, en este capítulo se aborda el caso de la evaluación multiparamétrica de la producción científica de las instituciones de educación superior (IES). En la sección 1.2 se definen diversos indicadores cienciométricos para medir la producción, la productividad y el impacto del quehacer científico de las instituiciones. En particular, se propone el ISP que mide la productividad y es aplicable a las instituciones mexicanas porque considera el número $ANnrs$ de miembros del SNI (Sistema Nacional de Investigadores del CONACyT) adscritos a cada institución. Además, se utilizan otros indicadores provistos por el SIR (Scimago Institutional Rank) como son: el impacto normalizado NI (eficacia), el $\%Q1$ (impacto esperado) y el $\%Exc$ (excelencia).

En principio suponemos que, estos parámetros son independientes y aunque puedan existir altas correlaciones entre pares, no existe una relación funcional determinista entre ellos. Esta suposición da lugar al difícil problema de realizar automáticamente una caracterización multiparamétrica de los perfiles de desempeño de las IES. Para hacer frente a este problema hemos ideado un método de minería de datos basado en la familia SOM. El método constituye una novedosa herramienta para el análisis y clasificación computacional multifactorial, presentada aquí como una aplicación de minería de datos cienciométricos. Se muestra la utilidad del método propuesto considerando los perfiles de desempeño cienciométrico de las 50 IES más productivas de México.

La propuesta metodológica involucra: el diseño y cómputo de indicadores que, considerarán distintas fuentes de información, el diseño y aplicación de métodos para la normalización de datos, así como la adecuación de métodos de clustering y visualización con redes SOM con capa enrome (ESOM).

En las siguientes secciones se mostrarán las modificaciones a los métodos SOM estandar facilitan la identificación y caracterización de perfiles de desempeño no-triviales, la identificación de datos anómalos (outliers) y el descubrimiento de correlaciones. En las aplicaciones reportadas en las publicaciones derivadas de este trabajo, se detallan las ventajas analíticas de la propuesta metodológica desde la perspectiva del experto del dominio de conocimiento que se está minando. En la sección 4.1 se presentan los datos necesarios para el estudio de las IES mexicanas. En la sección 4.2 se justifica el uso de un SOM con una capa de salida de gran tamaño y se demuestra gráficamente la mejoría que se consigue haciendo una transfomación adecuada de los datos. En la sección 4.3 se interpretan los resultados obtenidos en la caracterización multiparamétrica de los perfiles de desempeño en la

producción científica de las instituciones de educación superior mexicanas. Finalmente, en la sección 4.4 se discuten los resultados obtenidos desde la perspectiva del análisis cienciométrico.

4.1. Análisis multiparamétrico del desempeño de instituciones mexicanas de educación superior

De acuerdo con el Ranking Iberoamericano de Universidades 2013 (SCImago Research Group 2013), las 50 IES seleccionadas han producido más de 150 documentos, cubiertos por Scopus, durante un período de cinco años $(2007-2011)$. Dado que, la mayoría de los científicos productivos mexicanos están afiliados a las IES, el sistema de educación superior es el sector de producción de investigación más importante de México; de un total de 357 IES inscritas en el SIR en 2011, las 50 más productivas con el 80,21 % del total de los miembros del SNI.

En este estudio se presenta un análisis comparativo de la dinámica de la producción científica en las 50 IES mexicanas con mayor producción en Scopus en el período 2007-2011. Para esta comparación, se utilizan distintos recursos de visualización de datos y finalmente se utiliza una metodología basada en redes neuronales artificiales desarrollada para identificar automáticamente los principales perfiles de desempeño cienciométrico dentro de este grupo de IES. El análisis considerando conjuntamente todos los indicadores, es la principal ventaja de la metodología propuesta.

La metodología que aquí se presenta es aplicable para hacer estudios cienciométricos a varios niveles: micro (investigadores), macro (países) o a un nivel intermedio (meso). En este capítulo se presenta un estudio a nivel meso, analizando la producción de las instituciones mexicanas de educación superior. El análisis en cada uno de estos niveles implica minar un conjunto de datos $X \subset \mathbb{R}^n$, donde cada $x \in X$ es la representación de una unidad productiva, la cual queda determinada por una batería de indicadores cienciométricos que acusan distintos aspectos cuantitativos de su producción, su calidad y su impacto. Los indicadores cienciométricos utilizados están definidos en la sección 1.2.

En la sección 4.1.1 se plantea el problema de analizar tablas de indicadores cienciométricos; en la 4.1.2 se realiza un análisis de los datos considerando cada uno de los indicadores individualmente y en la 4.1.3 se analizan algunos diagramas de dispersión que resultan de graficar pares de variables. Las preguntas que deja sin resolver tanto el análsis unidimensional, como el análisis

biparamétrico motivan el uso de la red neuronal SOM.

4.1.1. Perfiles de desempeño cienciométrico

Como ya hemos establecido, por perfil de desempeño cienciométrico entendemos la representación de una IES como un vector $x \in \mathbb{R}^n$. Al considerar varias IES, el conjunto de perfiles de desempeño se puede mostrar como una matriz de datos. En la tabla de la figura 14 se muestran los valores de 10 indicadores calculados para las 50 IES mexicanas con más artículos publicados en revistas indexadas por *Scopus*. En esta tabla se pueden apreciar los valores obtenidos del cómputo de indicadores de volumen de producción ($Ndoc$, $\%Ndoc, ANdoc$), recursos de producción ($Nnrs$, $\%Nnrs$), productividad (ISP, IPR), impacto (NI, $\%Exc$) e impacto esperado ($\%Q_1$).

Para ejemplificar la técnica de análisis multiparamétrico se eligieron los siguientes cuatro indicadores: ISP (Productividad Científica Institucional), NI (Impacto Normalizado), $\%Q_1$ (Porcentaje en Cuartil 1), $\%Exc$ (Porcentaje de Excelencia). Ante una colección de datos como los que provee la tabla 14 uno se puede formular las siguientes preguntas:

- ¿Como se relacionan los rankeos inducidos por cada uno de los indicadores?
- ¿Cuáles son los patrones de desempeño que exhiben?
- ¿Cómo se relacionan los distintos patrones de desempeño?
- ¿Existen datos atípicos además de los que presentan valores extremos?

Mostraremos ahora varias maneras de obtener información y conocimiento a partir de los datos contenidos en la tabla de la figura 14.

4.1.2. Análisis y visualización uniparamétrica

Una primera forma de realizar el análisis de la información contenida en este conjunto de datos, es partiendo de la matriz tal y como se muestran en la tabla de la figura 14. En la tabla que se presenta en la figura 14 en cada una de sus columnas están resaltados aquellos valores que superan el valor promedio. Con este recurso visual se puede verificar fácilmente que los ordenamientos que establece cada variable son distintos unos de otros. En general, esta representación visual no aporta elementos de análisis o patrones que sean fáciles de interpretar.

Otra alternativa de presentación visual de los datos se muestra en la figura 15, donde se exhibe la magnitud de cada variable usando gráficas de barras.

Higher Education Institution	Ndoc	%Ndoc	Nnrs	%Nnrs	ANdoc	ISP	IPR	NI	%Exc	%Q1
Univ. Nac. Auton. de Mexico (UNAM)	19349	28.83	3320	21.78	3869.8	1.17	1.32	0.79	7.31	45.46
Cent. de Invest. y de Est. Avanz. (CINVESTAV)	7072	10.54	621.6	4.08	1414.4	2.28	2.58	1.03	9.68	42.05
Inst. Politec. Nac. (IPN)	5581	8.32	668.2	4.38	1116.2	1.67	1.9	0.63	5.27	30.69
Univ. Auton. Metropolitana (UAM)	3934	5.86	816.6	5.36	786.8	0.96	1.09	0.68	6.54	34.88
Univ. de Guadalajara (UdG)	2097	3.12	579	3.8	419.4	0.72	0.82	0.53	3.56	32.81
Univ. Auton. de Nuevo Leon (UANL)	1884	2.81	348	2.28	376.8	1.08	1.23	0.62	5.62	29.62
Benem. Univ. Auton. de Puebla (BUAP)	1876	2.8	331.8	2.18	375.2	1.13	1.29	1.07	8.54	34.38
Inst. Tecnol. y de Est. Sup. de Monterrey (ITESM)	1649	2.46	236.4	1.55	329.8	1.4	1.59	0.82	8.36	29.23
Univ. de Guanajuato (UG)	1579	2.35	217.8	1.43	315.8	1.45	1.64	0.77	7.78	37.68
Univ. Mich. de San Nicolas de Hidalgo (UMSNH)	1488	2.22	262.8	1.72	297.6	1.13	1.29	0.88	7.46	33.74
Univ. Auton. de San Luis Potosi (UASLP)	1482	2.21	213.6	1.4	296.4	1.39	1.58	1.09	9.64	42.91
Univ. Auton. del Estado de Morelos (UAEM)	1302	1.94	207.2	1.36	260.4	1.26	1.43	0.79	6.38	42.63
Univ. Auton. de Baja California (UABC)	1198	1.78	191.4	1.26	239.6	1.25	1.42	0.59	4.89	25.38
Colegio de Postgraduados (COLPOST)	1130	1.68	226.4	1.49	226	1	1.13	0.34	2.23	16.81
Univ. Auton. del Estado de Mexico (UAEMEX)	1075	1.6	238.8	1.57	215	0.9	1.02	0.56	4.89	27.44
Univ. Veracruzana (UVER)	880	1.31	233.4	1.53	176	0.75	0.86	0.56	4.61	34.32
Univ. Auton. de Yucatan (UAY)	821	1.22	128.2	0.84	164.2	1.28	1.45	0.78	8.16	34.47
Univ. de Sonora (USON)	801	1.19	179.2	1.18	160.2	0.89	1.01	0.53	3.96	36.2
Univ. Auton. del Estado de Hidalgo (UAEH)	747	1.11	151.4	0.99	149.4	0.99	1.12	0.46	2.79	27.44
Univ. Auton. de Queretaro (UAQ)	676	1.01	118.8	0.78	135.2	1.14	1.3	0.71	6.41	33.73
Univ. Iberoamericana (UIBER)	578	0.86	94.6	0.62	115.6	1.22	1.39	2.42	24.4	59.69
Univ. de Colima (UCOL)	538	0.8	108.2	0.71	107.6	0.99	1.13	0.63	4.64	40.89
Univ. Auton. Chapingo (UACHAP)	521	0.78	108.2	0.71	104.2	0.96	1.1	0.36	2.04	15.93
Univ. de las Americas Puebla (UAMERP)	521	0.78	54.8	0.36	104.2	1.9	2.17	0.78	7.95	27.45
Univ. Auton. de Zacatecas (UAZAC)	442	0.66	110.4	0.72	88.4	0.8	0.91	0.56	4.5	29.41
Univ. Auton. de Sinaloa (UASIN)	429	0.64	112.2	0.74	85.8	0.76	0.87	1.53	12.5	40.79
Univ. Auton. de Tamaulipas (UATAM)	391	0.58	62.8	0.41	78.2	1.25	1.41	0.55	5.7	23.02
Inst. Tecnol. de Tijuana (ITTIJ)	379	0.56	16.8	0.11	75.8	4.51	5.08	0.74	9.73	12.66
Univ. Auton. de Ciudad Juarez (UACJ)	338	0.5	66.6	0.44	67.6	1.02	1.14	0.69	6.64	28.7
Univ. Auton. de Aguascalientes (UAAGU)	318	0.47	51.2	0.34	63.6	1.24	1.4	0.4	3.35	23.58
Inst. Tecnol. de Celaya (ITCEL)	317	0.47	35	0.23	63.4	1.81	2.05	0.9	10.3	35.33
Univ. Auton. de Chihuahua (UACHIHU)	291	0.43	46	0.3	58.2	1.27	1.43	0.52	5.93	26.12
Univ. Juarez Auton. de Tabasco (UJATAB)	279	0.42	47.4	0.31	55.8	1.18	1.35	0.38	1.57	20.79
Univ. Auton. Agraria Antonio Narro (UAAAN)	278	0.41	44	0.29	55.6	1.26	1.42	0.4	4.43	22.3
Univ. Juarez del Estado de Durango (UJED)	253	0.38	31	0.2	50.6	1.63	1.87	0.67	4.96	23.72
Inst. Tecnol. Auton. de Mexico (ITAM)	250	0.37	65.6	0.43	50	0.76	0.86	0.65	4.83	34.4
Cent. Nac. de Invest. y Desarr. Tecnol. (CNIDT)	244	0.36	21.6	0.14	48.8	2.26	2.54	0.77	5.96	17.62
Univ. Auton. de Tlaxcala (UTLAX)	243	0.36	47.2	0.31	48.6	1.03	1.16	1.11	10.5	39.09
Univ. Auton. de Coahuila (UACOAH)	217	0.32	47	0.31	43.4	0.92	1.04	0.66	5.21	28.57
Univ. Auton. de Campeche (UACAMP)	214	0.32	33.2	0.22	42.8	1.29	1.47	0.67	7.73	37.85
Univ. Auton. de Baja California Sur (UABCS)	205	0.31	28	0.18	41	1.46	1.69	0.57	2.56	31.71
Univ. Auton. de Guerrero (UAGUE)	205	0.31	32.4	0.21	41	1.27	1.46	1.14	13.7	31.71
Univ. Auton. de la Ciudad de Mexico (UACM)	199	0.3	59.6	0.39	39.8	0.67	0.77	0.59	3.14	41.71
Univ. Panamericana (UPAN)	198	0.3	37	0.24	39.6	1.07	1.24	0.47	4.94	21.72
Inst. Tecnol. de Morelia (ITMOR)	183	0.27	14	0.09	36.6	2.61	2.94	0.42	1.08	18.03
Univ. Tecnol. de la Mixteca (UTMIX)	180	0.27	19.6	0.13	36	1.84	2.1	0.69	5.52	13.89
Inst. Tecnol. de Toluca (ITTOL)	166	0.25	6.6	0.04	33.2	5.03	5.77	0.39	2.78	22.29
Inst. Tecnol. de Veracruz (ITVER)	159	0.24	18	0.12	31.8	1.77	2.03	0.41	1.31	15.72
Univ. Auton. de Nayarit (UANAY)	159	0.24	20.6	0.14	31.8	1.54	1.78	0.63	2.88	29.56
Univ. Auton. de Chiapas (UACHIA)	154	0.23	37.6	0.25	30.8	0.82	0.93	0.38	2.74	26.62
Averages:	1309	1.951	215.4	1.413	261.9	1.4	1.59	0.70	6.15	30.25

Figura 14: Tabla de datos cienciométricos de las 20 IES mexicanas que más artículos de investigación tienen registrados en Scopus durante le periodo 2008 – 2013. Los valores mayores que el promedio aparecen resaltados.

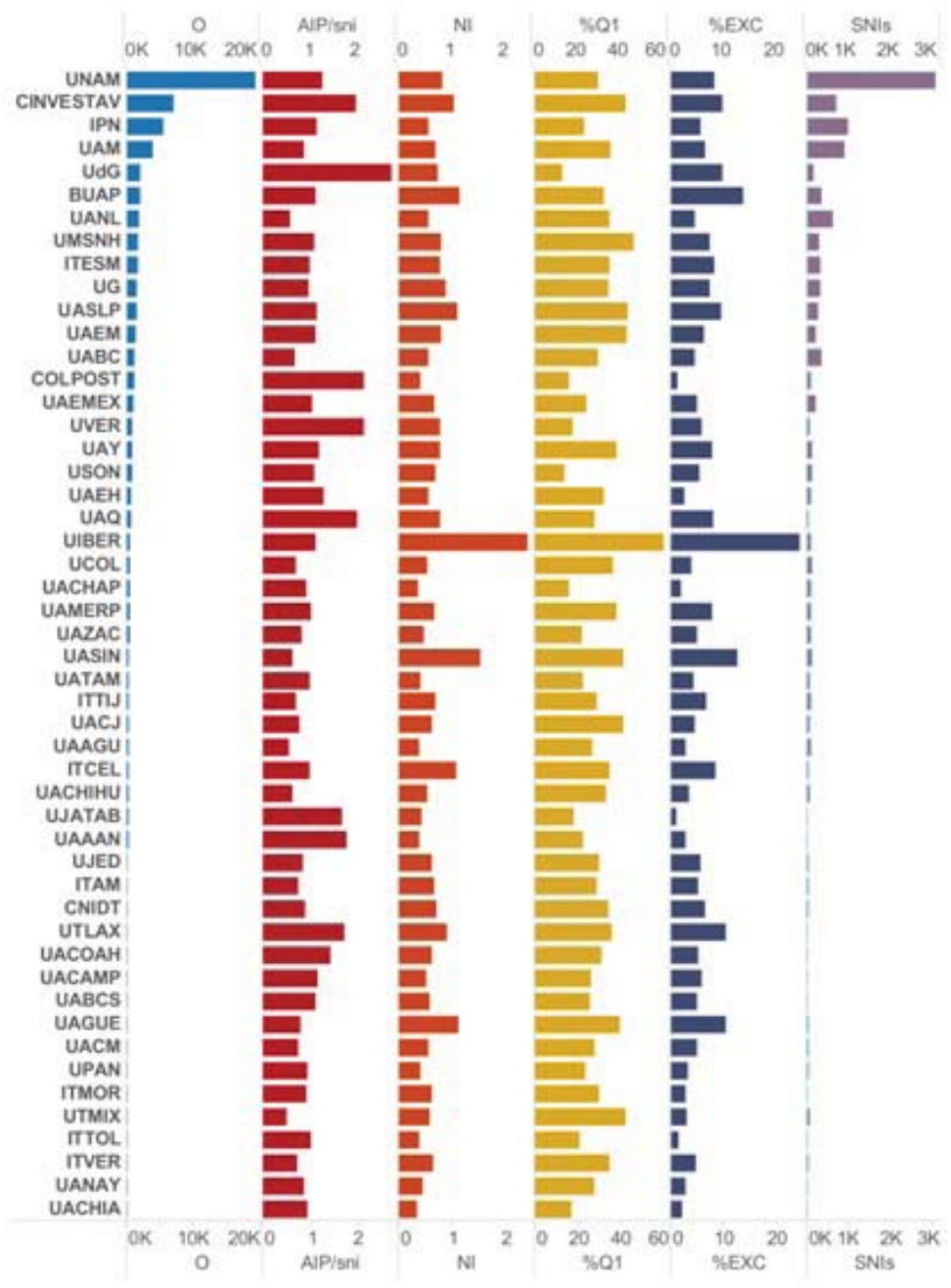


Figura 15: Visualización de datos de la tabla de indicadores cienciométricos asociados a las IES mexicanas más productivas durante el período 2008-2013.

A continuación se exponen algunos de los resultados obtenidos del análisis de los valores de cada uno de los indicadores desplegados en las tablas.

Productividad Institucional

La UNAM destaca con el mayor número de miembros del SNI y el mayor volumen de producción científica durante el período estudiado, produciendo 19,349 artículos, que representan el 28,83 % de la producción nacional y con el 21,78 % de los investigadores del SNI. Estos valores arrojan un cociente de productividad institucional de 1,32. Su producción promedio anual en este período fue de 3870 artículos, siendo este un valor extremo para este indicador.

Además de la UNAM, sólo hay otras dos instituciones que tuvieron una producción media anual superior a 1000 artículos: CINVESTAV (1414) y el IPN (1116). Estas instituciones concentraron 10,5 % y 8,3 % de la producción científica mexicana, con 4,08 % y 4,38 % del total de investigadores nacionales, obteniendo tasas de productividad institucional de 2,28 y 1,67, respectivamente. Con un IPR de 1,09, la Universidad Autónoma Metropolitana (UAM) es la segunda institución con más miembros del SNI (5,16 %) y la última con más del 5 % de la producción científica del país.

Durante el período $2007 - 2011$, el indicador NSP tiene un valor promedio de 0,86 y las 50 IES más productivas tienen un valor de ISP de 1,4. Por lo tanto, estas 50 instituciones tienen una contribución relevante a la productividad científica nacional. Por este indicador destacan, Instituto Tecnológico de Toluca (ITTOL, 5,03), el Instituto Tecnológico de Tijuana (ITTIJ, 4,51), Instituto Tecnológico de Morelia (ITMOR, $2,61$), el CINVESTAV (2,28) y el CNIDT (2,26). En el otro extremo, las IES con el ISP más bajo son: la Universidad Autónoma de Chiapas (UACHIA, 0,82), la Universidad Autónoma de Zacatecas (UAZAC, 0,8), el Instituto Tecnológico Autónomo de México (ITAM, 0,76), Universidad Veracruzana (UV, 0,75), Universidad de Guadalajara (UdG, 0,72) y la Universidad Autónoma de la Ciudad de México (UACM, 0,67).

Visibilidad, impacto y excelencia

Teniendo en cuenta la relación inversa de productividad e impacto observada a nivel macro, cabe esperar que, la eficiencia de producción de cada institución (número de documentos por miembro del SNI) no está correlacionada positivamente con la eficacia de la producción (el impacto de la producción, estimado en términos de número de citas).

La tabla de la figura 14 muestra el impacto normalizado (NI) de su

producción científica (comparado con el mundo), así como, el porcentaje de artículos publicados en las revistas indexadas por *Scopus* más visibles (Q_1), y $\%Exc$, el porcentaje de artículos que pertenecen al conjunto del 10 % de los artículos más citados de cada categoría temática de Scopus. En esta tabla siete IES mexicanas dentro de 50 más productivas tuvieron un mayor impacto normalizado que el promedio mundial durante el período $2007 - 2011$, y nueve IES tienen un $\%Q_1$ mayor que el promedio de las 50 IES seleccionadas (30,25 %). Las 50 instituciones tienen valores positivos del indicador de excelencia ($\%Exc$), lo que significa que, tenían al menos un artículo dentro del Núcleo de Excelencia de *Scopus*.

La Universidad Iberoamericana (UIBER) no se destaca en este grupo de instituciones por su productividad: con un $ISP = 1{,}22$ está por debajo del valor medio de las 50 IES. Sin embargo, es el más destacado en este grupo de IES, logrando valores extremadamente atípicos en los tres indicadores basados en citas: UIBER publicó el 59,7 % de sus artículos en revistas con un alta visibilidad, obteniendo un impacto normalizado de 2,42 (sus artćulos obtuvieron 142 % más citas que el promedio mundial). El 24,4 % de su producción científica se ubicó en el "núcleo de excelencia", que es una de las mejores marcas de las instituciones de América Latina, comparable a las puntuaciones de algunas de las más prestigiosas universidades de Estados Unidos y Reino Unido (Universidad de Harvard 28,69 %, Universidad de Stanford 27,29 % , y la Universidad de Cambridge 24,56 %). El número de autores en publicaciones UIBER también tiene valores muy extremos: el 45 % de sus artículos tienen más de 500 autores y estos artículos acumulan el 80 % del total de citas del período $2007 - 2011$. Estos datos podrían indicar que esta institución tiene un importante nivel de participación en grandes proyectos internacionales de investigación.

El valor promedio del NI fue de alrededor de 0,7 dentro del conjunto de las 50 IES más productivas. Hay seis IES además de UIBER, que han alcanzado este valor: la Universidad Autónoma de Sinaloa (UASIN, 1,53), Universidad Autónoma de Guerrero (UAGUE, 1,14), Universidad Autónoma de Tlaxcala (UTLAX; 1.11), Universidad Autónoma de San Luis Potosí (UASLP, 1,09), Benemérita Universidad Autónoma de Puebla (BUAP, 1,07) y CINVESTAV (1,03). Otras instituciones como el Colegio de Postgraduados (COLPOST, 0,34) y la Universidad Autónoma de Chapingo (UACHAP, 0,36) mostraron los valores más bajos en este indicador. Para fines de comparación a nivel mundial, es conveniente considerar las instituciones que tuvieron los más altos valores de impacto normalizado en 2011: Massachusetts Institute of Technology (29,86), The Rockefeller University (29,18), Harvard University (28,28) y London Business School (27,01).

El indicador de excelencia ($\%Exc$) mostró un valor promedio de 6,2 % en

el conjunto de las 50 IES, dentro de este conjunto 20 instituciones tuvieron este indicador por encima de la media. Las instituciones con los más altos valores son UAGUE (13, 7 %), UASIN (12,5 %), UTLAX (10,5 %) y el Instituto Tecnológico de Celaya (ITCEL, 10,3 %) son las instituciones con más del 10 % de sus trabajos en el núcleo de excelencia. Los valores más bajos del indicador de excelencia corresponden al Instituto Tecnológico de Morelia (ITMOR: 1,1 %) y al Instituto Tecnológico de Veracruz (ITVER: 1, 3 %). A nivel mundial en el ranking superior del indicador de excelencia tenemos los siguientes valores: London Business School (70,33), The Rockefeller University (61,64), Massachusetts Institute of Technology (57,87) y Harvard University (56, 84).

Con el 45, 5 % de sus trabajos publicados en las revistas de primer cuartil de Scopus, la UNAM fue la segunda en publicar en revistas de mayor visibilidad, después de UIBER 60 %. La Universidad Autónoma de San Luis Potosí (UASLP, 42,9 %), la Universidad Autónoma del Estado de Morelos (UAEM, 42,6 %), CINVESTAV (42 %) y la Universidad Autónoma de la Ciudad de México 41,71 % son los líderes de este indicador que tiene un valor medio del 30,2 %. Sin embargo, algunas instituciones registraron valores por debajo del 20 %. Notablemente, el ITTIJ, que tiene el quinto lugar de excelencia en el país, con $\%Q_1 = 12{,}7$ tiene el porcentaje más bajo de las 50 universidades estudiadas.

Conclusiones del análisis uniparamétrico

Ambos recursos de visualización y análisis (fig. 14 y fig. 15) se complementan, pero la visualización de la figura 15 permite identificar fácilmente valores extremos y algunas correlaciones entre los indicadores. Considerando el valor de los hallazgos en los análisis realizados, son evidentes sus limitaciones: el análisis unidimensional no permite determinar visualmente el grado de correlación que existe entre todas las parejas de indicadores. Cada uno de los indicadores establece un ordenamiento distinto y no es fácil comparar los patrones de desempeño de distintas instituciones.

4.1.3. Análisis biparamétrico

El análisis biparamétrico es un recurso común de visualización. El análisis de datos exploratorio permite identificar a "simple vista"la dispersión y relaciones funcionales entre dos variables.

Para ejemplificar el análisis biparamétrico, consideramos como variables de referencia al ISP y al $\%Exc$ y comparamos éstas con las demás variables. En particular, en la figura 16 se observa que, a excepción del CINVESTAV, IPN, ITESM y UG, los valores de ISP más altos (alta productividad) fueron

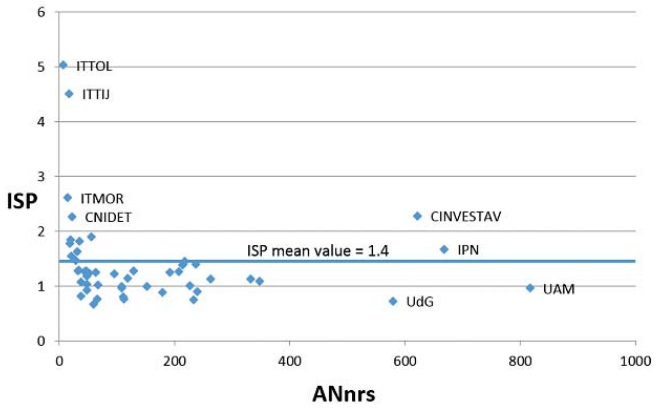


Figura 16: $ISP_{VS}Nnrs$ de las universidades mexicanas más prolíficas durante el período $2007 - 2011$.

obtenidos por IES con un número relativamente pequeño de investigadores nacionales ($Nnrs < 100$). Esto indica la complejidad intrínseca en la comparación de IES de tan diversas capacidades productivas.

En las figuras 17 y 18 presentamos los datos de estas 50 instituciones desde dos perspectivas diferentes: en primer lugar, detectamos instituciones que tienen el rendimiento bibliométrico más eficaz, trazando el impacto normalizado y la excelencia (fig. 17); en segundo lugar, comparamos eficacia versus efectividad, contrastando productividad con excelencia (fig. 18). La figura 18 complementa el análisis biparamétrico que compara la excelencia ($\%Exc$) con la visibilidad esperada ($\%Q_1$). En todas estas cifras excluimos a $UIBER$ para obtener una mejor imagen del comportamiento más típico. Este análisis abarca el período 2007-2011.

La figura 17 muestra el grupo de instituciones mexicanas con mejor desempeño cienciométrico desde una perspectiva basada en la eficacia. Se consideran indicadores que no toman en cuenta el volumen o la eficiencia de la producción, sino, el impacto. Las instituciones se organizan en torno a una línea recta en el gráfico (con coeficiente de correlación $R^2 = 0{,}9$, demuestra que existe un grado considerable de correlación entre estos dos indicadores). En una situación de correlación perfecta ($R^2 = 1$) los valores de NI aumentan en proporción directa a los valores de $\%Exc$. Sin embargo, localmente, como en la esquina inferior izquierda de la figura 17 se pueden observar configuraciones de puntos que acusarían a una correlación negativa, que contradice la tendencia global de los puntos descrita por la recta de regresión. Cinco instituciones se destacan con valores más altos en estas dos compomentes ($UASIN$, $UTLAX$, $UASLP$, $BUAP$ y $CINVESTAV$), y 16

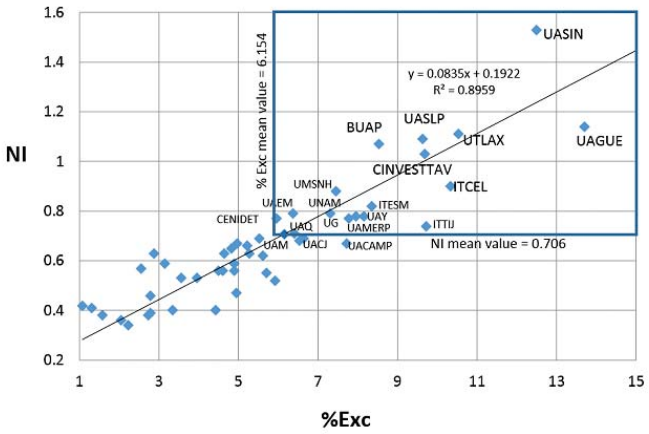


Figura 17: $NI_{VS}\,\%Esc$ en las IES mexicanas durante el periodos (2007-2011). La Universidad Iberoamericana no está representada en el cuadro porque es un valor extremo ($\%Exc = 24,41\,\%$), y un valor $NI = 2,42$. Para estos dos indicadores el coeficiente de correlación $R^2 = 0{,}9$.

instituciones se encuentran en el gráfico por encima de los valores medios de estos dos indicadores: $UIBER$, $UASIN$, $UAGUE$, $UTLAX$, $UASLP$, $UNAM$, $BUAP$, $CINVESTAV$, $ITCEL$, $ITTIJ$, $ITESM$, $UAMERP$, $UMSNH$, $UAEM$, UAQ y UAY.

En la figura 18 se observa una situación diferente, donde se exhibe la relación de excelencia ($\%Exc$) e indicadores ISP. No hay correlación entre estos indicadores (coeficiente de correlación $R^2 = 0{,}0002$) y las instituciones se distribuyen ampliamente en el plano cartesiano. Los valores más altos de $\%Exc$ fueron alcanzados por $UIBER$, $UAGUE$, $UASIN$, $UATLAX$; todos ellos tienen valores de productividad relativamente bajos (por debajo del promedio de las 50 IES estudiadas). El $ITTIJ$ es una notable excepción. Por encima de los valores promedio, y con un importante grado de equilibrio entre estos dos indicadores encontramos $CINVESTAV$, $ITCEL$, $UAMERP$ y UG. Por otro lado, aunque con una puntuación relativamente baja en excelencia, $ITTOL$ e $ITMOR$ destacan con valores muy altos del indicador de productividad. Pareciera que menos cantidad es el precio de la calidad, y viceversa.

Contrariamente a lo que cabría esperar, el gráfico de dispersión de la figura 19 muestra un grado muy bajo de correlación entre los indicadores (coeficiente de correlación $R^2 = 0{,}3755$). Sin embargo, un número importante de instituciones, 15 en total, obtienen puntuaciones por encima de los valores promedio en ambos indicadores. $ITTIJ$ muestra el comportamiento más

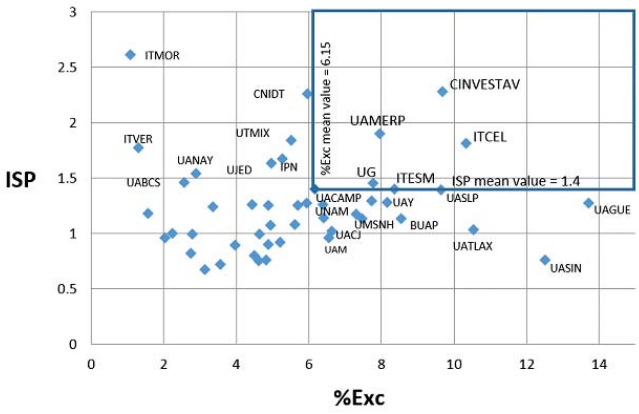


Figura 18: $\%Esc_{VS}ISP$ en IESs mexicanas durante (2007–2011). La UIBER no está representada en la imagen porque es un valor extremo de $\%Exc = 24,41\,\%$ y $ISP = 1,22$.

atípico: la puntuación más baja en $\%Q_1 = 12{,}6$ con puntaje muy alto en $\%Exc = 9{,}7$. Sin embargo, para el resto de las instituciones, el vacío del cuadrante inferior derecho (por debajo de $\%Q_1 = 30{,}25$ y hacia la derecha de $\%Exc = 6{,}15$) sugiere que el logro de la excelencia está fuertemente vinculado a la publicación en revistas de alta visibilidad. De manera diferente, el cuadrante superior izquierdo poblado por 8 instituciones sugiere que, con una probabilidad significativa, un pobre logro de excelencia puede ocurrir a pesar de publicarse en revistas de alto impacto o equivalentemente: "La baja excelencia no es el resultado de la publicación en revistas de bajo impacto".

Conclusiones del análisis biparamétrico

En el caso de estudio, la ausencia de relaciones lineales en los diagramas de dispersión sugiere que es importante la aplicación de métodos que consideren proyecciones no lineales.

La técnica de visualización biparamétrica podría generalizarse para analizar tres parámetros, pero con la dificultad que implica representar y visualizar puntos en el espacio tridimensional, razón por la cual, esto no se hace frecuentemente. Visualizar un diagrama de dispersión con más de tres parámetros es imposible, dada la imposibilidad humana de visualizar en cuatro o más dimensiones. Esta limitante conlleva a que solo sea posible descubrir patrones de comportamiento limitados a únicamente dos dimensiones explicativas.

Algunas de las cuestiones reveladas por el análisis biparamétrico se comprenden mejor desde una perspectiva multiparamétrica. En la siguiente sec-

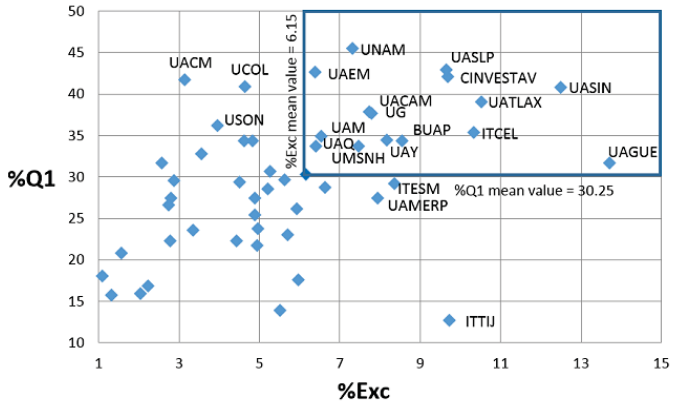


Figura 19: $\%Exc_{VS}\,\%Q_1$ para el periodo (2007–2011) con un coeficiente de correlación de $R^2 = 0{,}3755$.

ción un enfoque de inteligencia artificial, basado en técnicas de mapeo auto-organizadas, servirá para obtener una caracterización multiparamétrica y determinar los perfiles de desempeño cienciométrico de las universidades mexicanas más productivas.

4.2. Modelos SOM para el descubrimiento de patrones de desempeño

En este trabajo proponemos una técnica de minería de datos basada en la red neuronal SOM para el descubrimiento y caracterización de patrones de desempeño, en particular: la identificación de patrones raros (*outliers* multivariados) y el agrupamiento (*clustering*) de patrones similares. La dificultad para hacer este tipo de análisis se acentúa cuando el conjunto de datos es relativamente pequeño. Como es el caso del estudio que se presenta en este capítulo donde $X \subset \mathbb{R}^4$ y $\#X = 50$.

Se propone una variación del SOM original en la cual en primera instancia se transforman los datos. Después, en el entrenamiento se utiliza una red neuronal con un tamaño desproporcionado del número de neuronas frente al número de datos ($K >> N$). Para la visualización se modifica la propuesta original de los mapas de componentes (ver 3.1.1) para una interpretación más intuitiva de los mapas. A continuación se detallan y justifican estas modificaciones.

4.2.1. El efecto de la "normalización" de los datos en la visualización

En esta sección se detalla la normalización aplicada a los datos de las IES y se argumenta el porqué de su aplicación. En este caso los datos de entrenamiento que en un principio se consideran son de la forma $x \in \mathbb{R}^4$, donde

$$x = [IPR, \%Q_1, IN, \%Exc] = [x_1, x_2, x_3, x_4].$$

Las dimensiones tienen un rango de variación distinta porque $IPR, IN \in \mathbb{R}^+$, $\%Q_1$, $\%Exc \in [0, 100]$.

Para abolir el efecto diferenciado en los rangos de variación que las variables pueden tener sobre el cómputo de la distancia $d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}^+$ (*métrica euclidiana*) se realiza una normalización gaussiana, de manera que, cada $x \in X$,

$$\widehat{x}_i = \frac{x_i - \mu_{X_i}}{\sigma_{X_i}}. \tag{4.1}$$

Donde μ_{X_i} y σ_{X_i} son la media y la desviación estandar observadas en X.

El beneficio de esta normalización se observa claramente en los mapas de la Figura 20. En esta figura se disponen los mapas de componentes obtenidos al entrenar un Batch Map con distintos conjuntos de datos. A la izquierda están los mapas resultantes de entrenar al SOM con datos sin normalizar y a la derecha con datos normalizados.

Como se puede apreciar en el mapa de componente IPR existe una discontinuidad en el degradado de los colores. La esquina inferior derecha las tonalidad rojas se ubican en regiones disjuntas. Esta discontinuidad se debe a un "doblez"de la red en el espacio multidimensional. Los dobleces dificultan la interpretación y la hacen poco intuitiva. En los mapas de componentes obtenidos al entrenar la red con $\widehat{X}$ no se observan estos dobleces. La conlusión de este experimento es que la aplicación de la normalización evita dobleces en la red y los mapas resultantes tienen una interpretación mucho más clara.

4.2.2. Modelos SOM con capa enorme (ESOM)

En este trabajo proponemos el uso de modelos SOM con capa de salida $\mathcal{N}_O$ enorme, es decir, cuando el número de neuronas supera en órdenes de magnitud al número de datos de entrada $k >> N$.

En las aplicaciones del SOM que se reportan en la literatura es común que $N > k$; de hecho, uno de los usos que tiene el SOM es como cuantizador vectorial, para reducir el número de datos. Sin embargo, también es frecuente encontrar contextos de aplicación donde los datos no sean masivos. En este

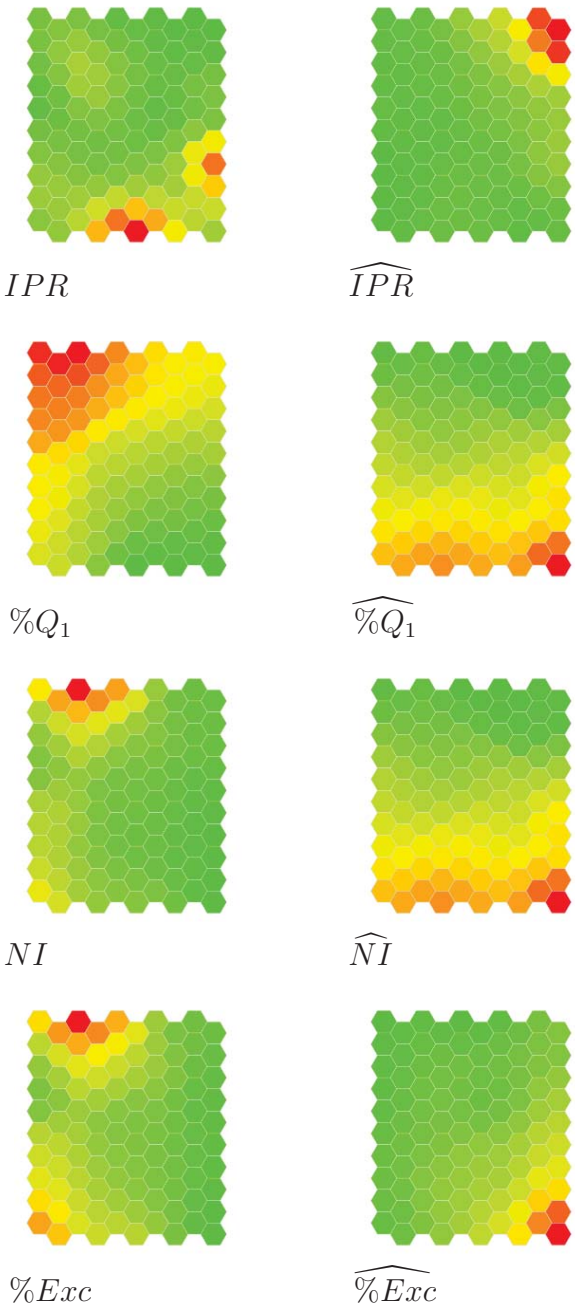


Figura 20: Comparación de Mapas de Componentes de X vs $\widehat{X}$

último caso tiene sentido plantearse la posibilidad de contar con una red neuronal tal que $K > N$. Además, se propone la razón entre el ancho y el alto de $\mathcal{N}_O$, corresponda con la razón entre los vectores principales de la matriz de covarianza. Ultsch y Herrmann prueban que, usando este tipo de redes es posible reducir el error topográfico de la red Ultsch and Herrmann [2005]. Recientemente, las redes ESOM se han utilizado en aplicaciones tales como: la caracterización de fenómenos complejos, la violencia doméstica Poelmans et al. [2010] y microarreglos Ultsch and Lötsch [2017].

Para obtener los mapas que se presentan en la sección 4.3 se utilizó un modelo SOM con una capa de salida $\mathcal{O}$ de tamaño (30×70), i.e. 2100 neuronas. Esta red se construyó para analizar un conjunto con apenas 50 datos.

4.3. Descubrimiento de patrones multiparamétricos de desempeño

La proyección ϕ provista por el SOM del conjunto $X \subset \mathbb{R}^4$, permite visualizar los datos de manera que las HEI pueden ser ubicadas en el mapa, asociadas a una región en el espacio multidimensional definido por los clusters. El supuesto más importante para hacer inferencias a partir del análisis de los mapas, es que la similaridad entre los desempeños de las distinta IES puede ser estimada al calcular la "distancia cienciométrica" entre las representaciones multidimensionales.

En este trabajo se aprovecha a la red neuronal para la no-trivial labor de identificar combinaciones raras que involucren a varios indicadores ($x \in \mathbb{R}^n$ con $n > 2$). Es muy simple identificar patrones raros si esto implica un valor extremo en alguno de los indicadores. Sin embargo, para la identificación de datos extremos, es conveniente usar una red neuronal con muchas más neuronas que datos ($K >> N$) contrario a lo que se observa en aplicaciones más típicas del SOM. Esta condición garantiza suficientes neuronas para los conjuntos de Voronoi con un solo dato.

4.3.1. Agrupamientos de perfiles de desempeño

Los mapas de clusters que se muestran en la figura 21 representan el conjunto de patrones de desempeño identificados por la red neuronal. Este mapa exhibe visualmente los datos extremos y los distintos estilos de desempeño.

El número óptimo de clusters se obtuvo mediante la aplicación del índice Dunn Dunn [1973]. En la siguiente gráfica se muestran los resultados obtenidos al aplicar el Índice de Dunn (ver figura 22).

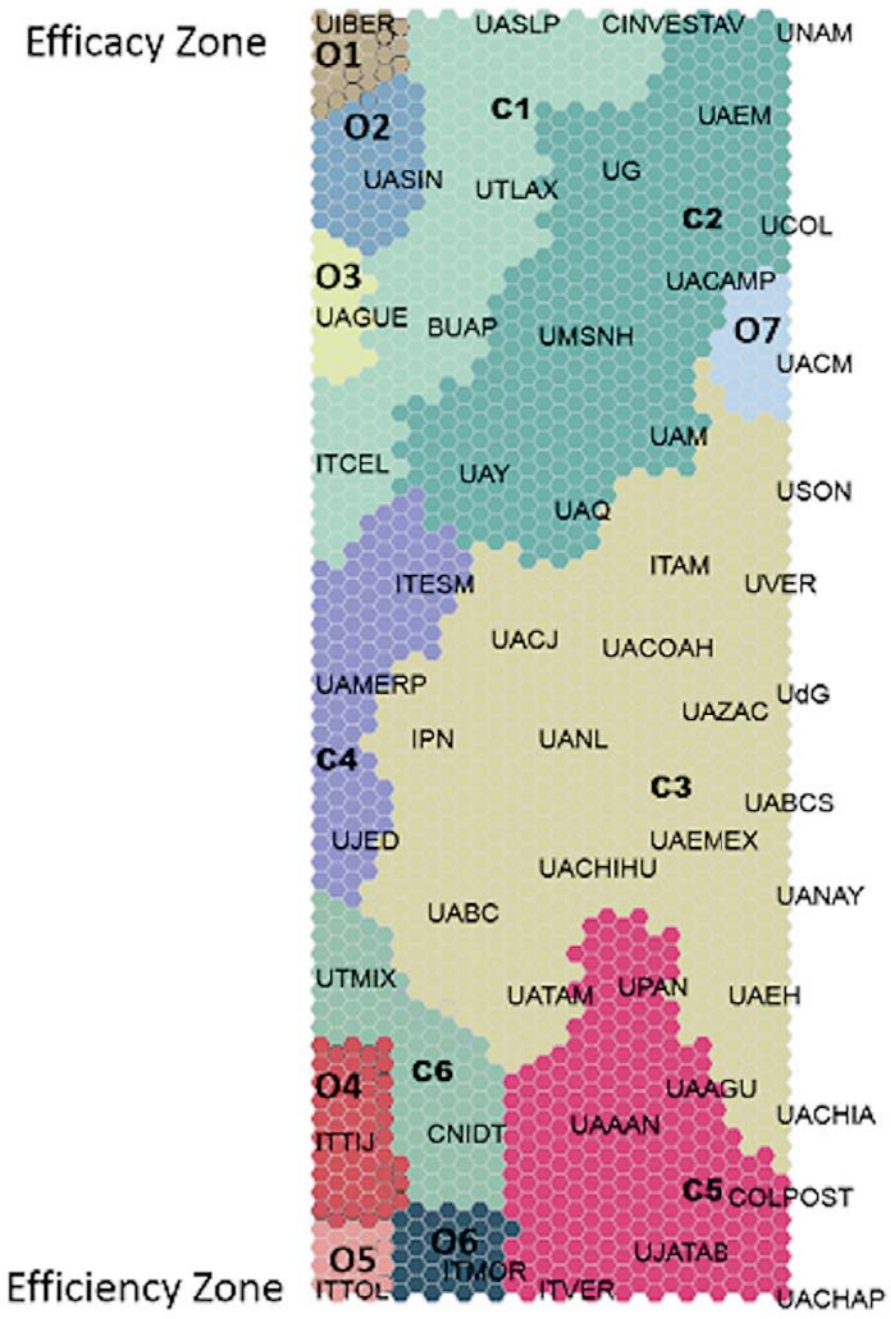


Figura 21: Mapa de Clusters de perfiles de desempeño de las IES

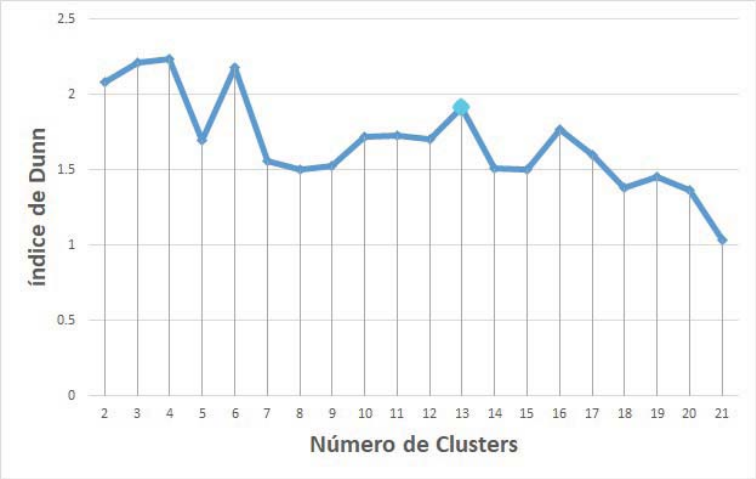


Figura 22: Gráfica de Índice de Dunn V_S Número de Cluster.

En el mapa de clusters (ver figura 21) se puede observar que la red neuronal identifica 13 clusters, siete de estos $O_1, ..., O_7$ se caracterizan por el hecho de que solo una institución se mapea en ellos; las otras seis regiones $C_1, ..., C_6$ contienen más de una institución.

4.3.2. Caracterización multiparamétrica

Para la interpreración de los resultados será de gran utilidad analizar los mapas de componentes de la figura 23. Se debe observar que, la parte superior del mapa de componente $\%Q_1$ corresponde a los valores más altos de este indicador. Conforme se aproxima a la parte inferior del mapa, este indicador decrece continuamente. Las instituciones que pertenecen a $O_1, O_2, O_3, O_7, C_1, C_2$ y la parte superior del cluster C_3 tienen los más altos valores de $\%Q_1$. A esta sección del mapa se le llamará *Zona de Alta Visibilidad Esperada*.

De igual manera, en el mapa de componente $\%Exc$, la parte superior izquierda corresponde a los valores más altos de este indicador, al igual que para el mapa de componente NI. Esto es de esperarse dada la alta correlación entre los dos indicadores. A esta parte del mapa se le llamará *Zona de Eficacia*. Las instituciones que tienen los más altos niveles en $\%Exc$ y NI se localizan en esta zona del mapa. Esta zona se intersecta con la Zona de Alta Visibilidad Esperada y los valores de $\%Exc$ y NI decrecen continuamente si se avanza en la diagonal que va a la esquina inferior derecha del mapa.

La esquina inferior derecha de todos los mapas de componentes es la

zona donde se presentan los niveles más bajos de los cuatro indicadores. La mayoría de las instituciones de C_5 se caracterizan por un perfil con una baja eficiencia, baja visibilidad y una baja eficacia.

Zona de Eficacia

Las instituciones agrupadas en esta zona tienen una buena cantidad de aplicaciones de alto impacto o participan en algún proyecto internacional de investigación que producen artículos (con cientos de autores) que acumulan una extraordinaria suma de citas.

El grupo elite de la Zona de Eficacia está compuesta por las instituciones: $UIBER, UASIN, UAGUE$ (O_1, O_2, O_3) y el grupo de instituciones que conforman el cluster $C_1 = \{UASLP, CINVESTAV, UTLAX, BUAP, ITCEL\}$. En esta misma zona, pero con menores valores en NI y $\%Exc$, se encuentra el cluster $C_2 = \{UNAM, UAEM, UG, UCOL, UMSNH, UAM, UAQ\}$.

La mayoría de las instituciones en la zona de eficacia tienen valores de ISP e IPR por debajo de la media de las top 50 instituciones de educación superior. El CINVESTAV tiene el perfil más balanceado entre las 50 instituciones más productivas. En general, las instituciones que pertenecen a la región C_2 se caracterizan por su alto porcentaje de publicaciones en el primer cuartil ($\%Q1$), los parámetros de eficacia ($\%Exc, NI$) y eficacia están por debajo del promedio.

Los tres perfiles extremos de la zona de eficacia O_1, O_2, O_3 tienen los más altos valores de $\%Exc$ y de NI, pero no sobresalen en sus valores de ISP. Tambien UASIN y UAGUE comparten este perfil de alta-eficacia/baja-eficiencia. UASIN es extremo en este sentido porque su valor ISP está cerca al mínimo observado entre las 50 universidades estudiadas, pero se encuentra entre las tres primeras en $\%Exc$ y NI. La red neuronal diferencía estos perfiles extremos al agruparlos en clusters distintos (O_2, O_3).

Zona de Eficiencia

En el mapa de componente ISP se observa en la esquina inferior izquierda, en la zona de eficiencia donde se localizan las instituciones con el más alto valor de ISP. En esta zona se pueden observar tres perfiles extremos O_4, O_5, O_6 y el cluster $C6$ se encuentra el grupo elite de la zona de eficacia. $ITTOL$ e $ITTIJ$ presentan los valores más altos de ISP del conjunto de 50 instituciones analizadas, seguidos por $ITMOR$ y $CNIDT$. Después de este grupo se encuentran $UTMIX$ e $ITVER$, con una menor eficiencia pero con un nivel de ISP por encima del valor promedio.

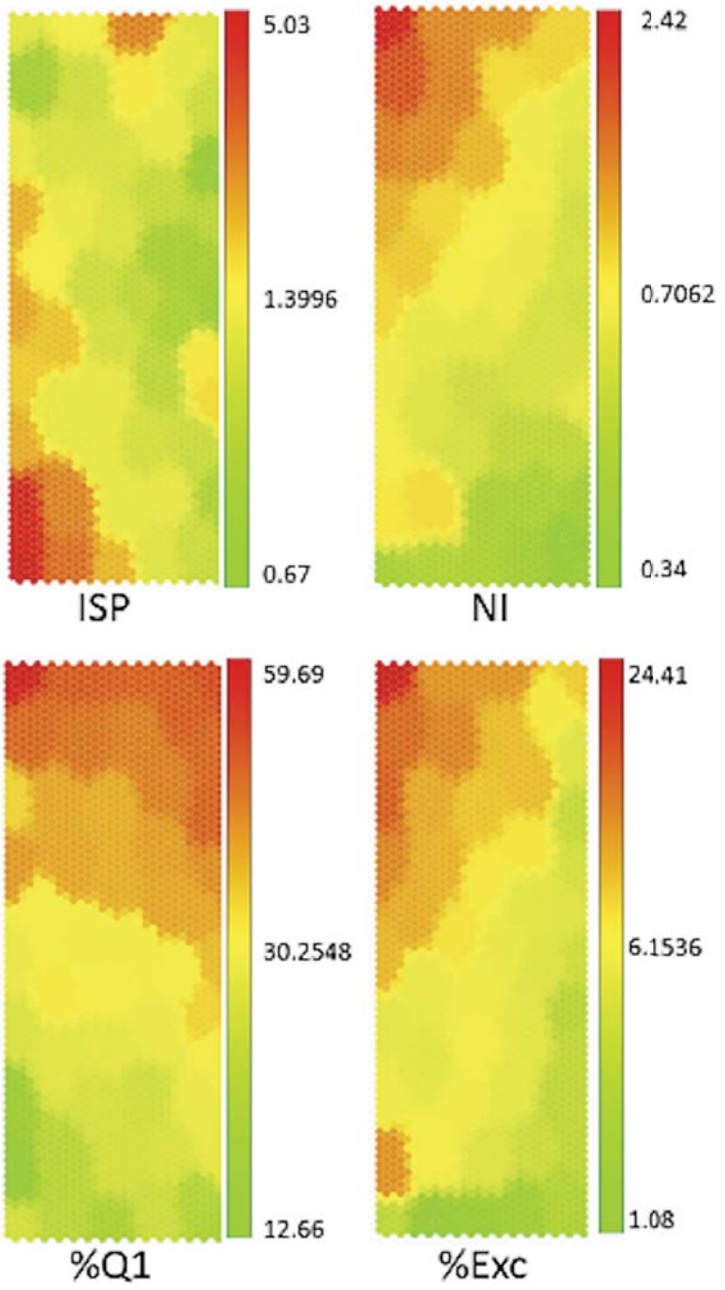


Figura 23: Mapas de Componentes de los indicadores de desempeño

Zona de Alta Visibilidad Esperada

En la parte superior del mapa de componente $\%Q_1$ se concentran instituciones que tienen los más altos niveles de porcentaje de sus plublicaciones en revistas dentro del primer cuartíl de las revistas más citadas. Como es de esperarse, la Zona de Eficacia es un subconjunto de la Zona de Visibilidad Esperada, lo cual significa que son las instituciones con valores altos de impacto. Un caso interesante excepcional es el de $ITTIJ$ (O_4): mientras que tiene un valor muy bajo en $\%Q_1$, $\%Exc$, y tiene el segundo lugar en productividad (ISP), estas características hacen que $ITTJI$ sea un dato extremo que no encaja en ningún perfil identificados.

Zona de Alta Visibilidad Esperada & Baja Eficiencia & Baja Eficacia

Existe una región triangular en la esquina superior izquierda del mapa donde se proyectan varias instituciones con valores relativamente altos de $\%Q_1$ pero con valores alrededor de la media en los indicadores ISP, $\%Exc$, NI. Este triángulo contiene la parte inferior de C_2, la parte superior de C_3 y O_7. Entre las instituciones con este perfil bibliométrico encontramos: $UACM$, UAM, UAQ, $USON$, $ITAM$ y $UVER$. El valor promedio de ISP de estas instituciones es bajo (1,4), pero, para algunas instituciones en particular es relativamente alto.

La existencia de esta Zona de Alta Visibilidad Esperada & Baja Eficacia refuerza el hecho de que publicaciones en revistas de alto impacto no necesariamente garantiza obtener un gran número de citas. Más aún, existe una situación en el cuál un alto impacto se alcanza sin la publicación en revistas de alto impacto. Esta situción se puede observar en la contraesquina de esta zona en la cual se ubica $ITIJ$, para la cual $\%Q_1$ tiene su valor mínimo.

4.4. Discusión de Resultados

El procedimiento de inteligencia artificial que hemos ideado ha sido aplicado con éxito para la minería de datos cienciométricos. Particularmente, fue útil realizar el análisis multiparamétrico y nos permitió identificar automáticamente diversos perfiles de desempeño cienciométrico institucional, las universidades que encajan en ellos, así como los perfiles de desempeño atípicos. Además, proporcionó recursos de visualización de datos claros y útiles que resultaron ser un excelente complemento al tradicional diagrama de dispersión y análisis de correlación. De todo esto concluimos que vale la

pena considerar el uso de este procedimiento de análisis y visualización para la minería de datos científicos y tecnológicos.

Las capacidades de proyección no lineal ofrecidas por la red neuronal SOM y la técnica de normalización de datos basada en la transformación logarítmica de una escala lineal por piezas apropiada, agrupa los datos y ofrece diferentes representaciones visuales que permiten inferir las relaciones de similitud entre los datos. El tradicional diagrama de dispersión y análisis de correlación es útil sólo para observar las correlaciones globales y los outlers se pueden identificar sólo en el caso donde hay un valor extremo. Además, las visualizaciones obtenidas son herramientas poderosas para descubrir valores atípicos que de otra manera serían muy difíciles de descubrir.

Se deduce de nuestro análisis la existencia de una correlación consistente entre la productividad (eficiencia de la producción) y la efectividad (entendida como el impacto de la producción) para las IES mexicanas (por ejemplo, las instituciones más productivas no son las que producen la investigación de mayor impacto). Este fenómeno tendría que ser investigado con más profundidad para relacionar la cantidad y la calidad de la producción científica de las IES mexicanas.

La batería de indicadores cienciométricos seleccionados no sólo considera el volumen de producción, sino que, también proporciona una estimación de la eficiencia y la calidad de la producción de investigación de las IES. Las diferencias de tamaño existentes entre las 50 instituciones de educación superior mexicanas más productivas, motivó el cálculo de nuevos indicadores independientes de tamaño para equilibrar las diferencias. Por ello, el uso de los datos del SNI fue fundamental.

Nuestro enfoque ha demostrado la importancia, no sólo de las instituciones más grandes del país, sino también algunas intermedias y, notablemente, pequeñas con puntuaciones sobresalientes en los indicadores cienciométricos. Esta es una indicación positiva de la existencia de excelentes investigadores en estas instituciones, lo cual es encomiable y digno de mención. No obstante, para fines de evaluación o para la comparación de las universidades más grandes con las más pequeñas, este hecho debe interpretarse con cuidado: el análisis estadístico, como el que hemos aplicado aquí, tiene en cuenta los promedios, y no es lo mismo promedio del trabajo de miles de investigadores (por ejemplo, el caso de la UNAM) que el promedio de sólo unas pocas docenas de ellos. En las grandes universidades "Ley de los Grandes Números" minimiza la posibilidad de irregularidades estadísticas. Además, las universidades grandes, a diferencia de los institutos tecnológicos relativamente más pequeños, tienden a ser temáticamente generalistas con grandes grupos de investigadores en las áreas de ciencias sociales, artes y humanidades, que no tienen los mismos patrones de publicación de ingenieros, matemáticos o

investigadores de ciencias naturales.

A pesar de la utilidad de la caracterización presentada en este estudio son necesarios más estudios para evaluar el desempeño de las IES y considerar otros aspectos de la producción científica. Por ejemplo, como un resultado colateral en nuestra investigación se evidenció la importancia de tomar en consideración el número de autores en publicaciones. Cabe destacar que la mayoría de los valores más altos en los indicadores $\%Exc$ e NI no corresponden a las IES mexicanas más grandes, sino a las relativamente pequeñas. Una situación similar está presente en relación con el indicador ISP. El indicador $\%Q1$ también muestra un comportamiento inesperado. Hemos encontrado que, con algunas excepciones, este indicador no se correlaciona positivamente con la productividad (indicador ISP) ni con los indicadores de impacto.

Capítulo 5

Descubrimiento de distribuciones (anti)simétricas en datos híbridos

El descubrimiento de asimetrías es un elemento de gran valor en escenarios como los estudios con enfoque de género. En estos casos los ejercicios analíticos se conducen con la lógica de develar diferencias en el desempeño escolar de hombres y mujeres.

En los estudios de género se asume que existe una variable categórica binaria (i.e. Sex), la cual determina dos clases $\{M, F\}$ donde:

$$M = \{x \in X, Sex(x) = 0\}, F = \{x \in X, Sex(x) = 1\}.$$

Normalmente esta variable es uniformemente distribuida, y para una población dada se observa que $\#F \approx \#M$.

Es relativamente sencillo verificar, a simple vista, la simetría (o asimetía) de la distribución de alguna variable dependiente Y respecto a la variable Sex. Por ejemplo, en la figura 24 se muestra una gráfica de barras de la distribución de estudiantes por estado de la república para cada una de las dos clases $\{M, F\}$. En esta gráfica se observa una distribución simétrica de la variable $Y = \#Estudiantes$ respecto a la variable independiente (factor)$X = Estado$ en las clases $\{M, F\}$.

A partir, de la identificación visual del patrón simétrico en la gráfica de la figura 24 se puede inferir que el vivir en un estado de la república en particular, no determina una diferencia de género en acceso a la educación. En este caso, el patrón geométrico de la gráfica acusa a una propiedad cuya verificación (o negación) es parte de los resultados esperados en un estudio de género. Normalmente los resultados del análisis de datos en este tipo de estudios se presenta contrastando $\{M, F\}$ con alguna variable dependiente

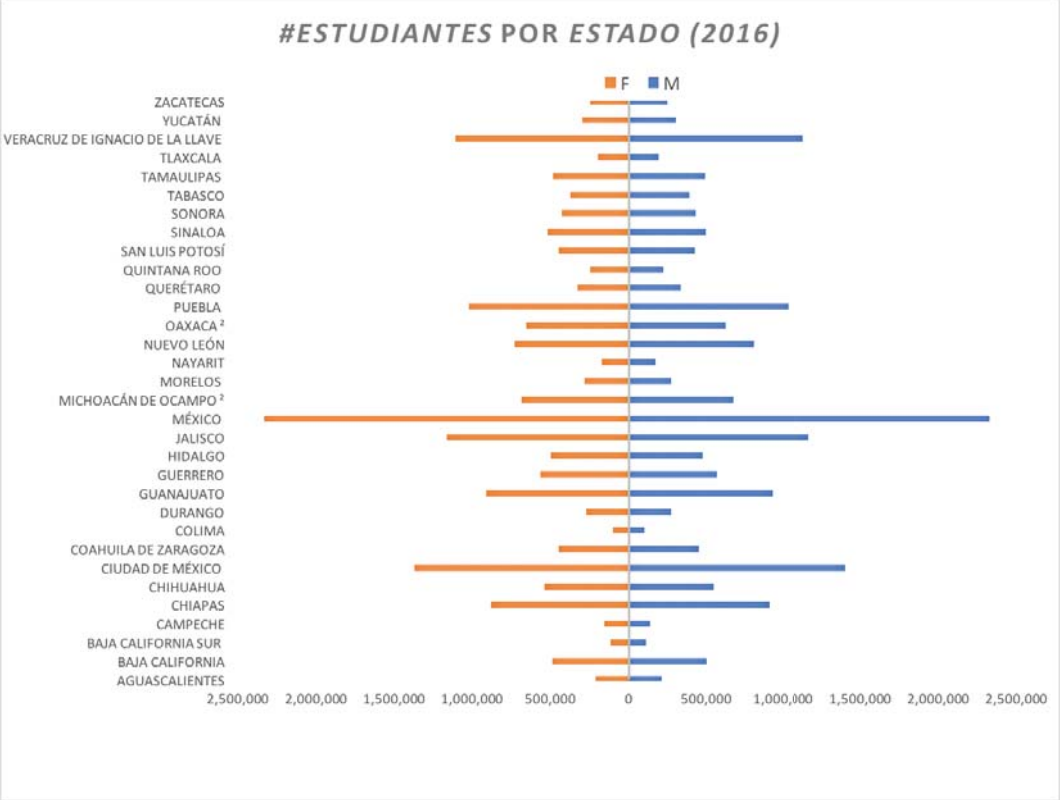


Figura 24: Simetría de la distribución por género de estudiantes por estado de la República Mexicana. Datos tomados de http://www.snie.sep.gob.mx/estadisticas_educativas.html

Y y cada uno de los factores X a analizar. Sin embargo, la identificación de diferencias de género que consideren múltiples factores $\{X_1, ..., X_n\}$ no es tarea trivial. En este capítulo se presentan aportaciones metodológicas en el uso de la red neuronal SOM, para el descubrimiento de diferencias de género multifactoriales.

En la sección 5.1 se describe la población estudiantil, se detallan las bases de datos que se usarán y se define el modelo matemático para representar a los estudiantes. El tipo de datos resultante de la modelación impone consideraciones especiales cuando se configura la red neuronal y al momento de interpretar los mapas. En la sección 5.2, se propone el uso de una métrica pesada que favorece la separación de las clases $\{M, F\}$ en la proyección sobre la retícula neuronal. Esta proyección evidencía los patrones simétricos o anti-simétricos en caso de que existan entre las dos clases. En la sección 5.3 se muestra el uso de los mapas generados por la red neuronal SOM. En su interpretación se aprovecha la capacidad de la visión humana para identificar patrones (anti)simétricos. Estos patrones geométricos se interpretan como posibles diferencias de género, las cuales puenden explicarse considerando múltiples factores.

El ejercicio analítico que se presenta utiliza información de estudiantes de nuestra la máxima casa de estudios, la Universidad Nacional Autónoma de México. Se consideran dos fuentes de información: el cuestionario de ingreso y el historial académico. Cabe señalar que, las dos bases de datos en cuestión son de uso reservado y el tener acceso a ellas significó contar con privilegios no convencionales. Los resultados obtenidos fueron reportados en publicaciones especializadas en estudios de género Villaseñor et al. [2008a], Millán et al. [2012]. Además, el Programa Universitario de Estudios de Género (PUEG) reconoció positivamente el valor del conocimiento obtenido.

La principal aportación de este estudio de caso es proponer una herramienta de análisis geométrico-cualitativo. Está ultima con la capacidad de representar de manera abstracta las asimetrías presentes, en las distribuciones de datos híbridos multidimensionales con respecto a los valores de una variable binaria Sex de control.

5.1. Modelación de la población estudiantil

El estudio de caso que se presenta en este capítulo se planteó originalmente como una contribución en la línea de investigación de universidad y género. En este caso se identificaron factores como son la formación familiar, la inserción laboral y la posición socioeconómica. Concebiblemente estos factores pueden afectar de manera diferenciada a las subpoblaciones femenina

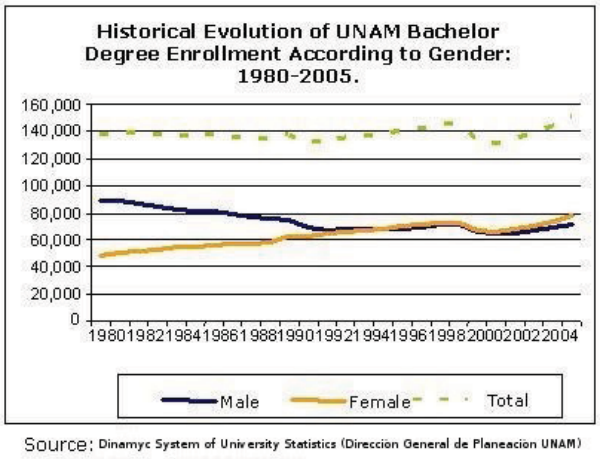


Figura 25: Proceso de feminización de la matrícula universitaria

F y masculina M en su desempeño como estudiantes en la universidad.

En la sección 5.1.1 se describe a grandes rasgos la población de estudiantes de la UNAM y cómo ha ido cambiando la composición en cuanto al número de mujeres y hombres, de pasar de una clara supremacía de la población masculina a un balance entre las dos poblaciones; en la sección 5.1.2 se describen las bases de datos a considerar en el estudio así como la modelación matemática que se lleva a cabo.

5.1.1. Proceso de feminización en la UNAM

Como se ilustra en la Figura 5.1.1, en el período de 1980 a 2005, la población estudiantil de la UNAM se equilibró en una cifra total de $140,000$ aproximadamente. Durante estos 25 años se ha observado un importante cambio en su composición por sexo; en 1980 la población de varones duplicaba a la de las mujeres, pero a partir de ese momento se observa un fenómeno de sustitución del espacio masculino por el femenino: la población femenina empieza a crecer notoriamente, mientras la masculina decrece en proporción complementaria.

Como consecuencia de esta tendencia sostenida, en 1994 las dos poblaciones se equipararon y ambas se mantuvieron con valores de alrededor de $70,000$ hasta el año 1999. En este momento una huelga que duró casi un año paralizó a la institución. Esto retrasó el ingreso de una generación teniendo el efecto de disminuir transitoriamente la población total. Después de esta caída numérica, ambas poblaciones se recuperan para alcanzar los valores que tenían antes de la huelga. Se produce entonces un interesante fenómeno:

la razón de recuperación de la población femenina supera notablemente al de la masculina.

Desde el año 2000 estas razones de cambio han mantenido su disparidad y para el 2005 la población masculina llega aproximadamente a los 73,000, mientras que, la femenina superó a su contraparte alcanzando los 80,000. Si se calcula la tasa de cambio que tuvieron estas poblaciones en el período 1980-2005, se observa un decrecimiento de la población masculina de más de un 25 % y un crecimiento de la femenina de más de un 40 %.

Este fenómeno de sustitución del espacio masculino por el femenino debe ser estudiado con mayor detenimiento, pues no es resultado de algún esfuerzo institucional especial por apoyar a este sector.

5.1.2. Modelación de población estudiantil como datos híbridos

Se consideraron los registros académicos de los alumnos de nivel licenciatura de la UNAM, campus Ciudad Universitaria, pertenecientes al sistema escolarizado de las cohortes de ingreso del período $1992 - 1996$ (ver Cuadro 5.1). La base de datos que se conformó para el análisis se obtuvo de dos fuentes: los historiales académicos y los cuestionarios de ingreso. Los datos de las dos fuentes fueron compaginados haciendo corresponder mediante los números de cuenta de cada estudiante. De estas cohortes se descartaron todos aquellos alumnos que no contestaron el cuestionario o no respondieron alguna de las preguntas que están involucradas en nuestro estudio. La muestra consolidada tiene un total de 39,893 alumnos lo que significa el 59,2 % de la población de las cinco cohortes estudiadas. Se hizo el seguimiento de estas cinco cohortes por un lapso de 20 semestres, que es el doble del tiempo establecido para cubrir el plan de estudios de la mayoría de las carreras. En la práctica después de este lapso el egreso reportado para todas las cohertes es prácticamente nulo.

De acuerdo a la tabla que aparece en el apéndice, la primera componente representa el sexo, considerado como variable binaria (masculino o femenino); la segunda componente es el área de conocimiento en la que está ubicada la carrera elegida; las cinco componentes siguientes tienen la información de las variables categóricas binarias, que en nuestra investigación constituyen los presuntos factores que pueden afectar el desempeño escolar: estado civil, hijos, empleo, escolaridad de la madre, posesión de automóvil en la familia; las siguientes veinte componentes tienen la información de los avances parciales (porcentaje de avance en su carrera). La última componente representa el indicador de egreso calculado a partir de las veinte componentes anteriores.

El indicador de egreso es una variable numérica discretizada para asumir cuatro valores correspondientes a las cuatro clases que determinan su estatus final, de acuerdo a la clasificación que se usa en la UNAM: egreso normativo, egreso límite, egreso terminal y no egreso. Estas cuatro clases están determinadas de acuerdo al tiempo que ocuparon los alumnos para cubrir un porcentaje igual o superior al 90 % de los cursos semestrales requeridos para terminar sus estudios de licenciatura. Los estudiantes incluidos en la clase "egreso normativo" corresponden a quienes acreditaron el 90 % de su plan de estudios durante los primeros 10 semestres. El grupo correspondiente a "egreso límite" es el que completó el 90 % entre el décimo y quinceavo semestre. El "egreso terminal" se determina cuando el 90 % de avance es sus créditos lo alcanzó entre el décimo quinto y vigésimo semestre. El "no egreso" cuando el alumno no cubrió el porcentaje después de 20 semestres.

La base de datos con la información de la población de alumnos se modela matemáticamente representando cada alumno por un vector en un espacio multidimensional (dimensión 28). Así, cada alumno está representado por 28 piezas de información numéricas. Cada una de estas piezas se corresponde con una componente del vector en el espacio multidimensional.

La variable Sex se modela como variable binaria que toma valores en $\mathbb{Z}_2$. Como se ha mencionado, esta variable determina las clases $\{M, F\}$ las cuales se usarán para comparar el efecto diferenciado de los distintos factores sociodemográficos en el egreso universitario.

Las variables dependientes corresponden a los diversos factores que acusan a condiciones sociodemográficas y que son modelados como variables categóricas binarias que toman valores en $\mathbb{Z}_2$ (Job, $MaritalStatus$, $Children$, $Automobile$) y ordinales $\mathbb{Z}$ ($Mother'sAcademicLevel$).

Como variable dependiente el indicador de egreso universitario Y es una variable ordinal que toma valores en $\mathbb{Z}$. Este indicador se obtiene del análisis de los avances parciales en cada semestre $\{Y_t\}_{t\in\{1,\ldots,20\}}$, como se explicó en la sección anterior.

Por simplicidad daremos el mismo tratamiento a las variables ordinales que a las reales. De manera que, el universo muestral en este caso tiene la forma

$$\mathbb{Z}_2^5 \times \mathbb{R}^{22}.$$

En este trabajo consideramos como datos con estructura híbrida (o simplemente híbridos) a aquellos que pueden ser modelados por un arreglo vectorial multidimensional. Existiendo componentes para representar variables numéricas y variables categóricas independientemente.

El análisis de datos híbridos es un problema que frecuentemente se presenta en el análisis informétrico. En estos casos, se consideran arreglos vectoriales

Variables	Descripción	Valores
Categóricas		
Sex	Sexo del estudiante.	$0 = Hombre$ $1 = Mujer$
Job	Tiene empleo al inicio de sus estudios.	$0 = No$ $1 = Si$
$MaritalStatus$	Está casado al inicio de sus estudios	$0 = No$ $1 = Si$
$Children$	Tiene hijos al inicio de sus estudios.	$0 = No$ $1 = Si$
$Automobile$	Su familia tiene auto al inicio de sus estudios.	$0 = No$ $1 = Si$
Ordinales		
$Mother's Academic Level$	Nivel de estudios de la madre.	$1 = Primaria$ $2 = Secundaria$ $3 = Preparatoria$ $4 = Universidad$ $5 = Posgrado$
$GraduationIndicator$	Tiempo que tomó concluir sus estudios.	$1 = Normativo$ $2 = Limite$ $3 = Terminal$ $4 = Avandono$
Numéricas		
$Avance_t$	Avance en el porcentaje de créditos en el mes t.	$[0, 1] \subset \mathbb{R}$

Cuadro 5.1: Descripción de variables consideradas en el estudio de género.

$x \in \mathbb{R}^n \times \mathbb{Z}_2^k$. Considerando el modelo presentado en (1.1), los datos híbridos corresponden a datos de la forma $n > 0$ y $k > 0$.

5.2. SOM con métrica pesada para tratamiento de datos híbridos

En la aplicación que se detalla en este capítulo, el uso de una métrica pesada como función de disimilitud en el criterio de competencia, juega un papel fundamental de interpretabilidad de los mapas obtenidos. Mediante esta métrica se modela matemáticamente la importancia relativa que las distintas variables tienen en el análisis y es clave para obtener representaciones visuales adecuadas. El diseño de la métrica pesada permite resaltar a la variable de control, Sex y promover que el ordenamiento global de los datos se haga de manera que las clases $\{M, F\}$ queden completamente separadas bajo la proyección.

La población de estudiantes es analizada contrastando los valores de factores socio-demográficos independientes; reflejados en variables categóricas binarias: Job, $Marital_status$, $Children$, $Automovil$ y la variable ordinal $MothersAcademicLevel$. Esta información se complementó con la provista por los avances semestrales de los alumnos en la matricula $\{Y_t\}$, reflejados en la evolución de su historial académico y el indicador de egreso Y.

Al entrenar la red neuronal utilizando esta métrica se representan visualmente las diferencias de género y quedan representadas como asimetrías en los respectivos mapas de componentes. De manera que, estos mapas evidencian cómo los distintos factores sociodemogáficos afectan de manera diferenciada el desempeño académico de hombres y a mujeres.

A continuación, en la sección 5.2.1 se define lo que se entiende por métrica pesada. En la sección 5.2.2 se discute el efecto que esta métrica tiene en el entrenamiento cuado esta se utiliza en datos de la forma $\mathbb{R} \times \mathbb{Z}_2^k$.

5.2.1. Métrica pesada y variables categóricas

Para cada $x, y \in \mathbb{R}^n$ con variables $x = (x_1, ..., x_n)$ y $y = (y_1, ..., y_n)$ la métrica pesada entre x y y está dada por

$$d(x, y) = \sqrt[2]{\sum_{k=1}^{n} (w_i(x_i - y_i))^2},$$

donde $w = (w_1, ..., w_n)$ es el vector de pesos asociados a cada dimensión. El uso de una métrica pesada durante el entrenamiento, modifica la proyección,

φ, del conjunto de datos, $X \subset \mathbb{R}^n$, a la malla bidimensional, $\mathcal{N}$.

Si suponemos que, todas las variables tienen el mismo rango de variación, el efecto de la métrica pesada sobre la proyección, se encuentra globalmente determinada por aquellas dimensiones con mayores y menores pesos. A continación se presentan y discuten algunos resultados sobre la aplicación de esta métrica utilizando datos con estructura híbrida.

5.2.2. Efecto de métrica pesada en datos híbridos

Consideremos una nube de puntos definidos sintéticamente, de la forma $X \subset \mathbb{Z}_2^2 \times \mathbb{R}$; es decir con una configuración de la forma $k = 2, n = 1$ (considerando (1.1)), donde cada variable es definida aleatoriamente con distribuciones uniformes para las tres variables, que dos a dos son aleatoriamente independientes.

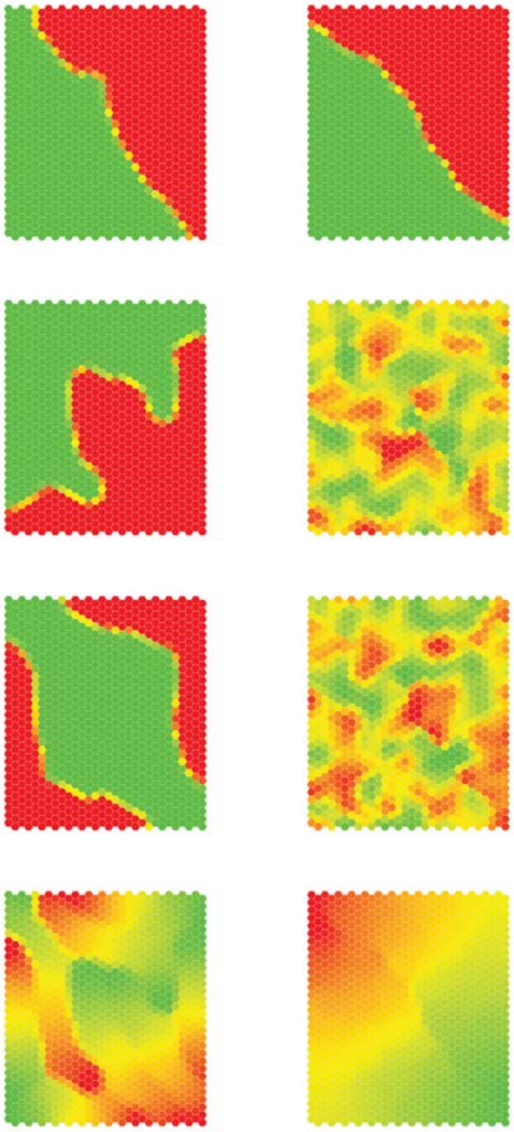
Este es un caso de dato híbrido muy simple pero es suficiente para definir los conceptos a tratar. En la siguiente sección se mostrarán los resultados de la aplicación de este tipo de métricas para el descubrimiento de diferencias de género impresas en el egreso universitario.

Como se discutirá a continuación, el efecto de las métricas pesadas sobre la proyección de los datos, implicará un recurso analítico que aporta el uso de estos modelos para el análisis de patrones asimétricos, en términos de imagen directa de conjuntos característicos de la forma $X_i = c$, con X_i variable binaria y $c \in \{0, 1\}$.

En la figura 26 se muestran los *mapas de componentes* resultado del entrenamiento de distintos modelos y configuraciones SOM. Como se puede apreciar, conforme se aumenta el peso de la variable categórica X_1 y se reduce el peso de X_2, la distribución de la variable numérica X_3 tiende a ser simétrica respecto a la diagonal, que es frontera entre las imágenes de los conjuntos $X_1 = 1$ (rojo) y $X_1 = 0$ (verde).

La siguiente afirmación es una hipótesis plausible a partir de lo que se observa en la figura 26.

Hipótesis de Proyección Simétrica (HPS) Sea $X \subset \mathbb{Z}_2^k \times \mathbb{R}$ con variables aleatoriamente independientes y con distribución uniforme, y X_3 con valores uniformemente distribuidos en el intervalo $[0, 1]$. Y sea un modelo SOMwWM definido de manera que $(w_1 = 2, \{w_j \to 0\}_{j=2}^k, w_3 = 1)$ entonces, el patrón cromático del mapa componente X_1 es uniforme y simétrico, el patrón cromático del mapa componente $X_j, j = 2 : k$ es crispado y uniforme y el mapa componente X_3 es simétrico y con un patrón cromático difuminado que sigue el orden de la barra cromática.



(a) (b)

Figura 26: Comparación de mapas de componentes obtenidos con (a) el SOM Básico v_S (b) SOMwWM $w_1 = 4, w_2 = 0{,}1, w_3 = 1$.

5.3. Descubrimiento de diferencias de género

Villseñor y otros autores presentan el análisis de los obtenidos desde la perspectiva de los estudios de género Villaseñor et al. [2008a], Millán et al. [2012]. En particular, se buscaba caracterizar cómo influye la condición de género al desempeño académico, y cómo distintos factores sociodemográficos afectan de manera diferenciada el desempeño escolar de hombres y mujeres, siguiendo cinco generaciones de la UNAM. La aplicación consiste en ejecutar el algoritmo SOM para atacar el problema de la identificación de factores socio-económicos que afectan el desempeño académico de manera diferenciada a hombres y mujeres. Un primer planteamiento del problema, así como, resultados preliminares se encuentran reportados en el estudio de Villaseñor y colaboradores Villaseñor et al. [2008a]. Posteriormente, se recibió la invitación de publicar una versión extendida de este trabajo en el Journal of Resources for Femenist Research Millán et al. [2012]; revista internacional especializada en estudios de género.

El resultado final es una aplicación del algoritmo WMSOM en donde se confronta el problema de analizar datos híbridos: compuestos por variables numéricas, variables categóricas y variables ordinales. El planteamiento metodológico de la aplicación, así como, la interpretación de los resultados obtenidos se realizaron considerando aspectos metodológicos propios de la demografía y los estudios de género. Los detalles más técnicos de esta aplicación se detallan en la investigación de Villaseñor Villaseñor et al. [2008a].

En este estudio se supone que nubes de datos híbridos multidimensionales representan poblaciones de seres humanos y se considera la variable SEX como la variable de control, se plantea la siguiente pregunta:
¿Asimetrías en la distribución de variables asociadas a factores socio-demográficos son indicadores visuales de desigualdades, en el sentido de los estudios con enfoque de genero?

En el Figura 27 se distinguen, principalmente, el color rojo (mujeres) y azul (hombres). Las dos regiones son bastante simétricas y claramente diferenciadas por una curva diagonal que las divide. Estas regiones se denominan región femenina y región masculina, respectivamente. Las dos regiones son muy similares, sin embargo, la femenina es ligeramente más alargada. Esto es consistente con la ligera diferencia entre la población de estudiantes hombres y mujeres.

En la Figura 28(a) *Graduation Efficiency* se distinguen cuatro regiones: egreso normativo (rojo), egreso en tiempo límite (amarillo), egreso en tiempo terminal (verde agua), abandono (azul rey). Dado que la distribución de estas regiones es casi simétrica respecto a la linea diagonal que separa la población

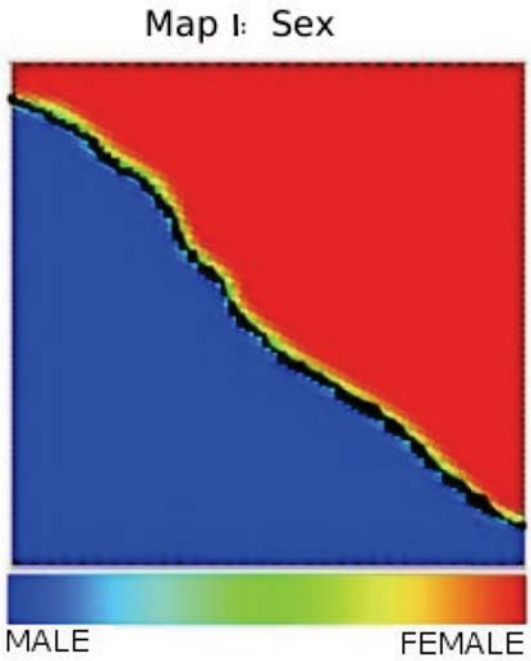
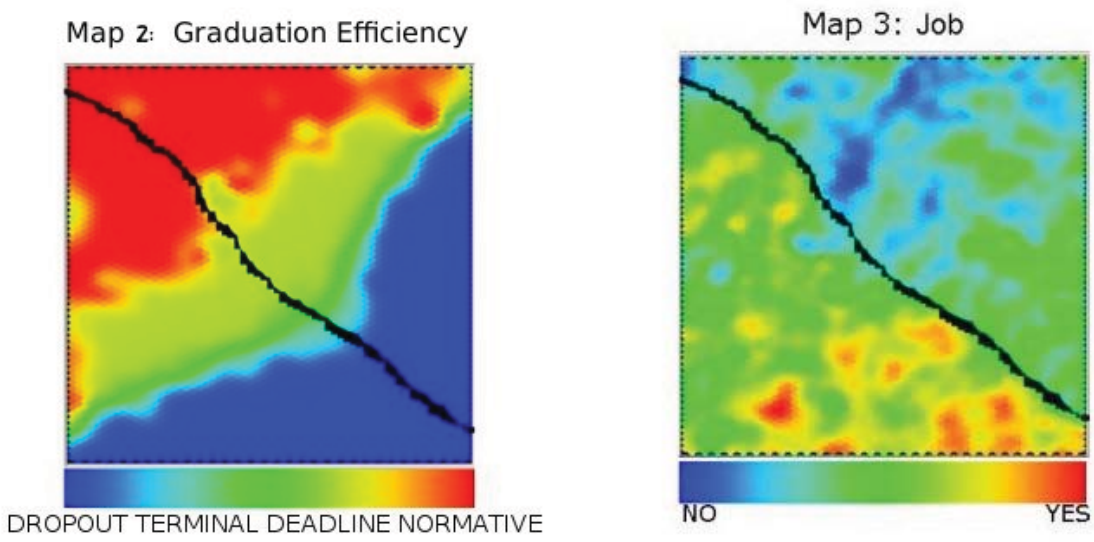


Figura 27: Mapa de componente SEX.

femenina de la masculina, este mapa revela que no hay diferencias significativas en la eficiencia terminal. Sin embargo, esta simetría menos marcada en la región de egreso normativo, el color rojo ocupa una área ligeramente más grande en la zona femenina. En este caso, la proporción de mujeres graduadas en tiempo normativo es mayor que la de los hombres.

La Figura 28(b) exhibe una marcada asimetría: las tonalidades que van del amarillo al rojo parecen predominar en la zona masculina, esto es indicativo de un alta presencia de estudiantes hombres que tienen empleo al inicio de sus estudios. Mientras tanto, en la zona femenina no se observan rojos pero si predomina el color verde. La baja proporción de mujeres que declaran estar trabajando al inicio de su carrera se puede explicar por el hecho de que el trabajo no remunerado (como labores domésticas) no está socialemente considerado como un trabajo.

También es interesante observar que (tanto para hombres como para mujeres) la mayoría de los casos de estudiantes que comienzan sus estudios con un trabajo se encuentran distribuidos en la región de abandono. Por lo tanto, el trabajo a una edad temprana que se puede asociar con el rol de proveedor (masculino) compíte con el rol de estudiante, y en muchos casos afecta el desempeño académico. Es importante mencionar que, la edad promedio para



(a) (b)

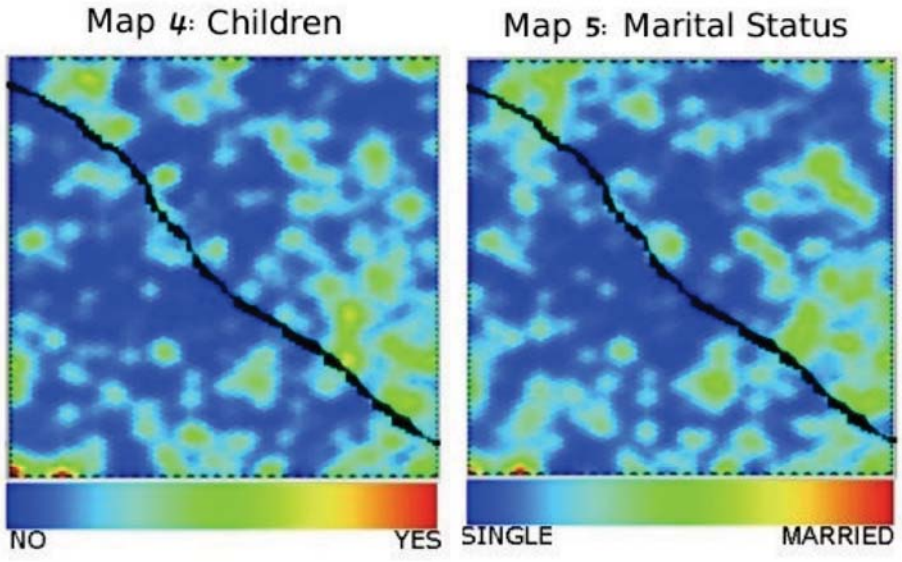
Figura 28: Mapas de Componentes *Graduation Efficiency* y *Job*

el comienzo de los estudios universitarios en la UNAM es de 19,97 años y en el mercado laboral mexicano no es común encontrar ofertas de medio tiempo, las cuales son convenientes para los estudiantes.

La Figura 29(a) muestra la distribución de los estudiantes que tenían hijos al comienzo de sus estudios universitarios. Aunque las tonalidades del amarillo al verde se observan en todo el mapa, estas tienen mayor presencia en la zona femenina, en la región de abandono, con algunos casos excepcionales en la zona femenina de egreso normativo.

Como es de esperarse, una situación similar se presenta en la Figura 29(b), el cual corresponde al estado marital. Existe una tendencia diferenciada, con respecto al sexo. Para aquellos estudiantes casados o con hijos al inicio de sus estudios, se correlaciona claramente con lo roles tradicionales de género, de acuerdo con los cuales es un deber de las mujeres hacerse cargo de la familia y el hogar, mientras que, las aspiraciones profesionales son para los hombres.

En las mapas de la Figura 30 se observan patrones similares. Las manchas rojas en el mapa de la Figura 30(a) de la corresponden a estudiantes cuyas madres tienen el más alto nivel de estudios (licenciatura y posgrado), mientras que, las manchas azules están con aquellos estudiantes con madres que solo terminaron la primaria. Por otro lado, las manchas rojas en el mapa



(a) (b)

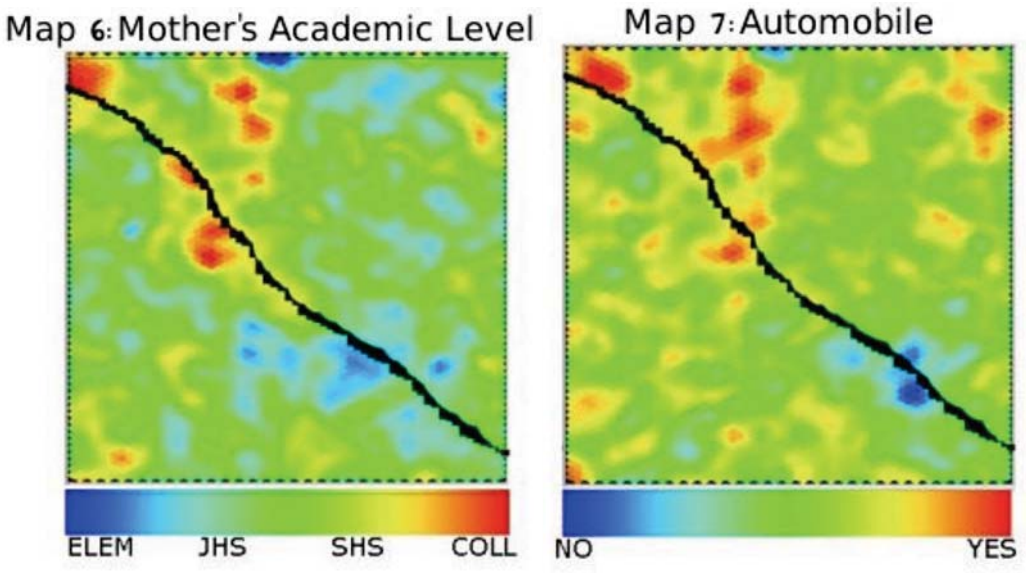
Figura 29: Mapas de Componentes *Children* y *Marital Status*

de la Figura 30(b) corresponden a estudiantes cuyas familias tienen un automóvil y las manchas azules a estudiante cuyas familias no tienen autómovil. La similitud entre estos dos mapas se correlaciona bien con la expectativa de que familias cuya madre tienen una mejor instrucción también tienen una situación económica más favorable.

Es de esperarse que un nivel socio-económico elevado de la familia implique una ventaja competitiva para el desarrollo económico. Al buscar asimetrías en los mapas de las Figuras 30(a) y 30(b), con respecto a la diagonal, se pueden descubrir efectos diferenciados de factores económicos en el desempeño escolar de hombres y mujeres. En ambos mapas las concentraciones más grandes de tonalidades rojas-amarillas se encuentran en la región de egreso normativo y las manchas azules en la región de abandono, lo anterior dentro de la zona femenina. A partir de este análisis se puede inferir que estos dos factores son importantes en el desempeño escolar de las estudiantes.

En la zona masculina no se observa lo mismo, la distribución de los colores es muy uniforme y no hay concentraciones notorias ni de rojo, ni de azul. Esto significa que, para los hombres estos factores no afectan de manera significativa su desempeño escolar.

A pesar de que la identificación de las causas de estas asimetrías va más allá del alcance de esta investigación, se pude establecer como conjeturas las



(a) (b)

Figura 30: Mapas de Componentes *Mother's Academic Level* y *Automovile*

diferencias de género impresas en el egreso universitario.

Además de los descubrimientos mencionados, como se dicutió en el capítulo 4, una de las ventajas de esta técnica es que permite la identificación de comportamientos anómalos. Por ejemplo, existe una pequeña mancha en los mapas de la Figura 30(a) y 30(b) que revela un grupo relativamente pequeño de estudiantes mujeres, en la parte superior de la región de egreso normativo, cuya familia tiene una situación económica desfavorable así como, una madre con baja instrucción. Además, sobreponiendo el mapa de la Figura 28(b) también se puede observar como algunas de estas estudiantes inclusive trabajaban al entrar a la universidad.

5.4. Conclusiones desde la perspectiva de estudios de género

Los resultados computacionales obtenidos por la metodología planteada son consistentes con la información disponible y los resultados de investigaciones previas, lo cual valida nuestro procedimiento. Con esta metodología se evidencían diferencias entre hombres y mujeres, sean estas ubicadas dentro del ámbito académico como es el caso o cualquier otro escenario en el que

se desee explorar, por su capacidad de auto-organización particiona de modo natural los mapas mostrando visualmente las diferencias.

En la figura 27 se observa una casi perfecta simetría que muestran las zonas femenina y masculina. Si alguno de los factores de desempeño propuestos no tuvieran un efecto diferenciado por sexo, se esperaría que las coloraciones de los correspondientes mapas de componentes se distribuyeran de manera simétrica respecto a la diagonal entre la zona femenina y la masculina.

Con esta investigación experimental se ha confirmando que la formación familiar, la inserción laboral y la posición socio-económica del estudiante son factores que afectan el desempeño escolar. Al considerar la Figura 28, que corresponde al indicador de egreso y contrastarlo con cada uno de los mapas componentes asociados a factores sociodemográficos, encontramos una variedad cromática que señala asociaciones entre el redimiento acadadémico y los factores considerados. Se observa que los factores afectan el desempeño del estudiante de manera diferenciada entre las clases $\{M, F\}$.

Por la ubicación e intensidad de la coloración de los mapas componenetes de formación familiar ($Children$,$MaritalStatus$) e inserción laboral (Job), se concluye que, estos factores afectan a cada uno de los sexos de manera diferenciada al desempeño académico. Ni el total de hombres, ni el total de mujeres son afectados idénticamente como subpoblaciones. En este caso, si se puede hablar de un grupo de hombres que es afectado negativamente por la inserción laboral asociado a su carrera. Al mismo tiempo, existe un grupo de mujeres que es afectado negativamente por la formación familiar en el rol de amas de casa y madres.

Aunque estas diferencias son visibles, es importante considerar la información en la que se basa la investigación. Esta última es obtenida de los cuestionarios en el momento de ingreso a la licenciatura de los estudiantes y no existe un seguimiento posterior, mientras que, ambos factores son típicamente cambiantes en el tiempo,específicamente, en el rango de edades en que se encuentra la población de estudio.

Adicionalmente, el análisis exploratorio de datos llevado a cabo permite hacer las siguientes predicciones que deben ser objeto de futuras investigaciones:

- La posición económica y el nivel de estudios de la madre tienen un peso importante para el egreso en tiempo normativo
- El nivel económico familiar tiene mayor impacto que la escolaridad de la madre para el egreso en tiempos normativos en la población masculina.
- El que la madre de un alumno tenga un alto nivel de escolaridad tienen un notable efecto positivo sobre la trayectoria académica de las

estudiantes, y no tanto así sobre los estudiantes.

- El efecto diferenciado que la escolaridad de la madre tiene sobre las alumnas, aunado al actual proceso de sustitución de los espacios masculinos por los femeninos, sugiere que esta tendencia de feminización se acenturá en el largo plazo. Esta conjetura debe ser objeto de futuras investigaciones.

El empleo de la metodología propuesta ofrece las siguientes ventajas:

- Provee representaciones visuales de la información que son fáciles de entender y comunicar.
- Permite identificar grupos minoritarios en la población. Usando otras técnicas de análisis estadístico, las características de los grupos minoritarios, podrían resultar no significativas respecto a las medias globales.
- Los resultados obtenidos pueden complementarse con técnicas del análisis estadístico convencional.

Capítulo 6

Visualización de dominios de conocimiento

En el año 2005 Chaomei Chen, editor en jefe de la revista *Information Visualization*, publicó el artículo Chen [2005] donde se establecen los 10 principales problemas no resueltos en el campo de la visualización de información. Dentro de esta lista de problemas se refiere el problema de *Visualización de Dominios de Conocimiento* (KDViz). Recientemente en Lu et al. [2017] se establece una lista de 15 problemas no resueltos dentro de los cuales el KDViz se ratifica como uno de los problemas de frontera, afirmando que este problema "*is a synthesized challenge which requires conveying of information structures with knowledge*".

En este capítulo se presentan aportaciones metodológicas desarrolladas para abordar la compleja tarea de analizar y visualizar dominios de conocimiento. Las aportaciones se basan en el uso de la red neuronal SOM para analizar datos multidimensionales. Estos datos resultan del cómputo de indicadores bibliométricos obtenidos a partir del procesamiento de conjuntos de registros bibliográficos. Uno de los elementos de mayor originalidad de esta propuesta es el análisis de secuencias temporales construidas a partir de la ocurrencia de descriptores, con la finalidad de identificar distintas fases en la evolución de dominios de conocimiento.

Nuestro método de analítico se ejemplifica dentro del campo Biomédico, analizando la producción mundial de investigación en el dominio de Vacunas de Tuberculosis (*Tb Vaccines*). Una primera versión de este estudio de caso se llevó a cabo en colaboración con un grupo de expertos en vacunas del *Instituto Finlay* de la Habana Cuba. Los resultados obtenidos se publicaron en el capítulo "*Bioinfometric studies in TB vaccines research*" del libro *The Art and Science of Tuberculosis Vaccines*, publicado por la editorial Oxford y que es referido en este trabajo como Guzmán et al. [2010]. Cabe señalar

que este libro fue merecedor del *Premio Anual de la Salud 2011* otorgado por el Ministerio de Salud de Cuba. En esta tesis se profundiza y se extiende la metodología utilizada para demostrar la utilidad del método propuesto para analizar y visualizar la evolución de dominios de conocimiento.

En la sección 6.1 se presentan algunas de las características que tienen los registros bibliográficos de dominios biomédicos recuperados de MedLine. En particular se resalta la importancia de contar con un conjunto de descriptores que se encuentran ordenados jerarquicamente en la ontología MeSH (*Medical Subject Heading*). Además, se presenta una de las aportaciones de este proyecto doctoral que consiste un método que permite aprovechar estos descriptores y el conocimiento experto de indexadores de MedLine, para caracterizar semánticamente dominios biomédicos.

En la sección 6.2 se discute la utilidad y las dificultades de analizar secuencias temporales producidas por los patrones de ocurrencia de descriptores con el algoritmo SOM. De los resultados obtenidos en una primera instancia se puede concluir que es necesario inventar una medida de similitud apropiada para las secuencias temporales. La propuesta es una medida de disimilitud que no cumple la desigualdad del triángulo.

En la sección 6.3 se muestran los resultados de analizar el dominio de conocimiento *TB Vaccine* utilizando la medida de similitud temporal introducida. El resultado final es un mapeo temporal de la red de coocurrencia de descriptores MeSH, asociada a este dominio de conocimiento.

Finalmente en la sección 6.4 se discuten las conclusiones desde la perspectiva de la visualización de dominios de conocimiento (KDViz) relacionadas con las aportaciones presentadas en este capítulo.

6.1. Dominio de conocimiento biomédico: MedLine

Desde el año de 1957 hasta el primer semestre del año 2017, la Biblioteca Nacional de Medicina (NLM) de Estados Unidos, ha indizado cerca de 27 millones de documentos que datan desde inicios del siglo XIX. Estos documentos están relacionados con temas de investigación en biomedicina (medicina, enfermería, odontología, oncología, medicina veterinaria, salud pública, ciencias preclínicas y otras áreas de las ciencias de la vida).

La base de datos que han construido se llama MedLine y está disponible a través de PubMed (`www.pubmed.com`), en el portal del *National Center for Biological Information*. Hoy en día MedLine constituye la base datos más importante en Biomedicina a nivel mundial.

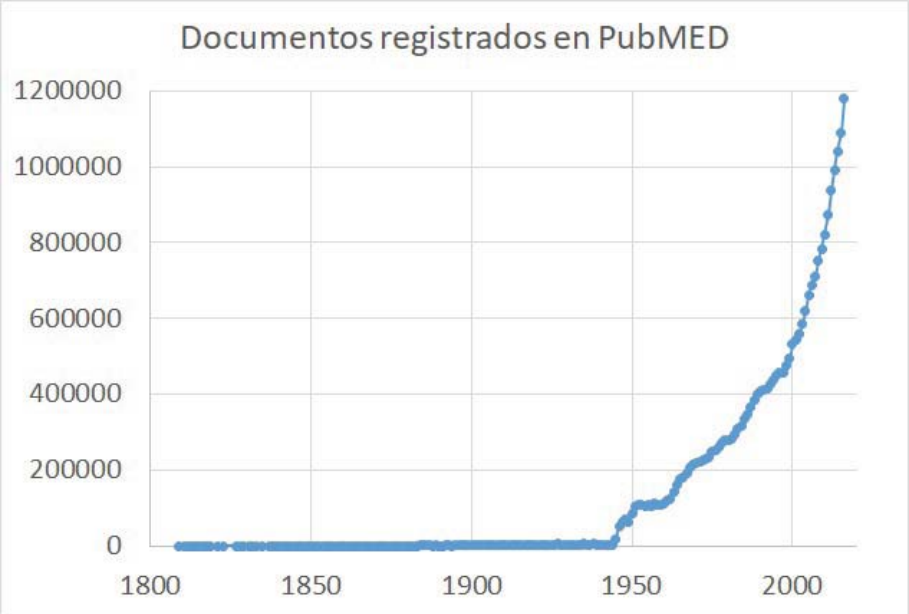


Figura 31: Producción científica anual en dominios biomédicos registrada en PubMED

Como se observa en la figura 31 la acumulación de documentos actualmente tiene un crecimiento cercano al exponencial. Cada año se publican más documentos y surgen nuevas revistas. Para facilitar el acceso y la categorización de esta gran colección de registros bibliográficos se han desarrollado diversas herramientas, en particular la ontología MeSH.

En la sección 6.1.1 se describe a grandes rasgos la estructura de la ontolgía MeSH y como se utiliza para describir el contenido de las publicaciones científicas. En la sección 6.1.2 se propone un método para determinar la relevancia de los términos MeSH basada en la coocurrencia no simétrica de descriptores. Finalmente, en la sección 6.1.3 se introduce el dominio "*Tb Vaccines*" desde la perspectiva informétrica. Este dominio será de utilidad para ilustrar los métodos analíticos que se proponen al final de este capítulo.

6.1.1. Ontología MeSH

La ontología MeSH organiza jerárquicamente los descriptores con los que se etiquetan las publicaciones. Cada uno de estos descriptores determina un dominio de investigación. De manera que la tarea de recuperar todos los documentos relacionados con una temática en particular se simplifica cuando esta temática queda determinada por un término MeSH.

Figura 32: Categorías principales del árbol MeSH y despliegue del primer nivel del ramal [L] Information Science.

El primer nivel de la ontología está compuesto por 16 categorías principales y cada categoría se ramifica en subcategorías cada vez más específicas. Esta ontología es revisada anualmente por un equipo de profesionales, con la finalidad de incorporar temas emergentes. La versión 2015 de la ontología contiene $27,455$ descriptores. En la figura 32 se muestran las categorías principales de esta ontología así como el despliegue del primer nivel del ramal que corresponde a *Information Science*.

Una de las ventajas que tiene el trabajar con la ontología MeSH es que la relevancia de las palabras clave se puede determinar aprovechando el conocimiento de un gran grupo multidisciplinario de expertos indexadores. Estos expertos son los encargados de etiquetar cada uno de los documentos contenidos en MedLine. Los indizadores utilizan marcas para determinar la relevancia que tiene un descriptor para una investigación en particular. Un término MeSH puede aparecer como: descriptor simple (con minúsculas), descriptor mayor (con mayúscula inicial), subtítulo (con el prefijo ":") o como un descriptor principal (con el prefijo "*"). En la siguiente sección se presenta un método que utiliza esta información para determinar la relevancia de cada descriptor en un dominio de conocimiento.

6.1.2. Cálculo de la relevancia de térmnios MeSH: centralidad y Google Pagerank

En general, las redes de palabras clave definida por una matriz de coocurrencia COO, como en 1.6, es simétrica. En este caso, consideramos que la matriz COO tienen ceros en la diagonal. Es decir, no consideramos las coocurrencias de un término en sí mismo. De manera natural, la matriz COO define una red pesada no-dirigida. La suma de los elementos en cada columna COO_j es el grado de centralidad del nodo j, es decir,

$$\deg(j) = \sum_i COO_{i,j}.$$

Gracias a la ponderación que realizan los indizadores de PubMed cuando etiquetan los artículos científicos con los términos MeSH, es posible establecer relaciones asimétricas. Consideramos una matriz COO^* con entradas proporcionadas por:

$$COO^*_{i,j} = \sharp\{d \in D \mid (\{kw_i, kw_j\} \subseteq kw(d)) \wedge (kw_j \text{ es el descriptor principal} r)\} \tag{6.1}$$

Una vez determinada la red dirigida inducida por la matriz de adyacencia COO^* es factible aplicar el algoritmo Pagerank Brin and Page [1998] para obtener un ordenamiento de los términos MeSH, el cual considera la relevancia de los descriptores en un dominio de investigación. Las premisas de este algoritmo son las siguientes Langville and Meyer [2004]:

- La relevancia de un nodo es la suma de las relevancias de los nodos que están conectados a ese nodo.
- Cuando un nodo relevante tiene muchas conexiones con otros nodos, su relevancia será proporcionalmente distribuida a todos los nodos a los que esté conectado.

Las columnas de la matriz COO^*_j se normalizan de tal manera que la suma de sus entradas es igual a 1, lo cual la hace una matriz estocástica. Estas columnas se utilizan para conformar una nueva matriz GOO,

$$GOO_j = \frac{1}{\sum_i^n COO^*_{i,j}} COO^*_j$$

La manera en que se define GOO, la mantiene como una matriz estocástica que define un proceso de Markov.

El algoritmo PageRank de Google consiste en aplicar el *método de potencia con un desplazamiento*, para determinar la distribución límite del proceso

de Markov que define la matriz estocástica GOO. Este método consiste en multiplicar la matriz(6.2) por si misma,

$$(1-\alpha)GOO + \alpha 1_{N\times N} \tag{6.2}$$

hasta que las columnas de la matriz resultante son numéricamente iguales. Aquí α es el parámetro de cambio entre 0 y 1, $1_{N\times N}$ es una matriz $N \times N$ llena de unos, y N es el número total de palabras clave.

En la tabla que se muestra en la figura 33 se muestran tres ordenamientos de los 50 descriptores más relevantes, que ocurren en el dominio "*Tb Vaccines*", de acuerdo a distintos criteríos de relevancia. En la primera columna con encabezado **GPR**, se muestra el ordenamiento inducido por el método propuesto en este apartado basada en el método *Google PageRank*. En la segunda columna **Grado de Centralidad**, se considera como criterio de ordenamiento la centralidad de los descriptores MeSH como nodos de el grafo inducido por la matriz COO. En la tercera columna **Frecuencia**, se muestra el ordenamiento dado por la ocurrencia total de cada término en el conjunto de documentos.

Como se puede apreciar, los criterios basados en centralidad (Google Pagerank, Grado de Centralidad) consideran como más relevantes a los descriptores que tienen mayor pertinencia en el dominio y no son tan generales. En la sección 6.2 se propone un método para analizar la evolución de la relevancia de términos MeSH. Este método está basado la caracterización de distintos momentos en la evolución de dominios biomédicos basado en el concepto de coocurrencia dirigida presentado en esta sección.

6.1.3. Dominio "*Tb Vaccines*"

Es bien sabido que la comunidad científica ha dedicado una considerable atención a las enfermedades respiratorias infecciosas. Considerando los registros bibliográficos de MedLine podemos observar que en particular la tuberculosis ha recibido una atención especial: la cantidad de artículos de investigación registrados desde 1950 en tuberculosis (131, 459 artículos) contrasta notablemente con el que corresponde a otras infecciones respiratorias, como la Pneumococcal (12,609 artículos), la Meningococcal (7,997 artículos) o la Legionellosis (4,024 artículos).

La figura 34 muestra la evolución del nivel de actividad de la investigación en tuberculosis y en ella se observa que esta ha sufrido importantes variaciones durante el período de estudio. A partir de un nivel de producción anual promedio de 3,438 artículos publicados durante los años 1950-1952, se nota un declive sostenido durante diez años (1952-62) hasta alcanzar un nivel

GPR	Grado de Centralidad	Frecuencia
BCG Vaccine	BCG Vaccine	BCG Vaccine
Tuberculosis,	Tuberculosis,	Humans
Urinary Bladder Neoplasms	Urinary Bladder Neoplasms	Female
Mycobacterium bovis	Tuberculosis, Pulmonary	Animals
Tuberculosis, Pulmonary	Mycobacterium bovis	Male
Mycobacterium tuberculosis	Tuberculin Test	Child
Adjuvants, Immunologic	Mycobacterium tuberculosis	Tuberculosis
Immunotherapy	Adjuvants, Immunologic	Adult
Tuberculosis Vaccines	Carcinoma, Transitional Cell	Mice
Tuberculin Test	Immunotherapy	Adolescent
Carcinoma, Transitional Cell	Tuberculosis Vaccines	Infant
Antigens, Bacterial	Vaccination	Child, Preschool
Vaccination	Antigens, Bacterial	Middle Aged
Neoplasms	Melanoma	Mycobacterium bovis
Macrophages	Leprosy	Urinary Bladder Neoplasms
Melanoma	Neoplasms	Aged
Bacterial Proteins	Hypersensitivity, Delayed	Tuberculosis, Pulmonary
Hypersensitivity, Delayed	Macrophages	Mycobacterium tuberculosis
Antineoplastic Agents	Carcinoma in Situ	Adjuvants, Immunologic
Leprosy	Tuberculin	Immunotherapy
Immunity, Cellular	Antineoplastic Agents	Vaccination
Carcinoma in Situ	Child	Infant, Newborn
T-Lymphocytes	Bacterial Proteins	Tuberculin Test
Tuberculin	Skin Neoplasms	Administration, Intravesical
Skin Neoplasms	Neoplasm Recurrence, Local	Guinea Pigs
Bacterial Vaccines	T-Lymphocytes	Antigens, Bacterial
Mycobacterium	Immunity, Cellular	Hypersensitivity, Delayed
Interferon-gamma	Lung Neoplasms	Carcinoma, Transitional Cell
Mycobacterium Infections,	Infant	Immunity, Cellular
Vaccines, DNA	Mycobacterium	Interferon-gamma
Immunization	Bacterial Vaccines	Time Factors
Granuloma	Mycobacterium Infections,	Macrophages
Lung Neoplasms	Antitubercular Agents	Follow-Up Studies
Neoplasm Recurrence, Local	Interferon-gamma	Antitubercular Agents
Vaccines	Immunization	Lymphocyte Activation
Lymphocytes	Granuloma	Mice, Inbred C57BL
Antitubercular Agents	Tuberculosis, Bovine	Immunization
Neoplasms, Experimental	Neoplasms, Experimental	Neoplasm Recurrence, Local
Tuberculosis, Bovine	Lymphocytes	Antineoplastic Agents
Antibody Formation	Isoniazid	Age Factors
Child	Vaccines, DNA	Tuberculosis Vaccines
Mycobacterium leprae	Vaccines	Antibody Formation
Lung	Research	Clinical Trials as Topic
Immunity	Antibody Formation	Mice, Inbred BALB C
Acyltransferases	Mycobacterium leprae	Melanoma
Isoniazid	Lymphadenitis	Lung
Research	Infant, Newborn	Bacterial Proteins
Antibodies, Bacterial	HIV Infections	Combined Modality Therapy
Vaccines, Synthetic	Breast Neoplasms	Lymphocytes
Antigens, Neoplasm	Communicable Disease Control	T-Lymphocytes

Figura 33: Ordenamientos distintos de los términos MeSH que ocurren en el dominio "*Tb Vaccines*".

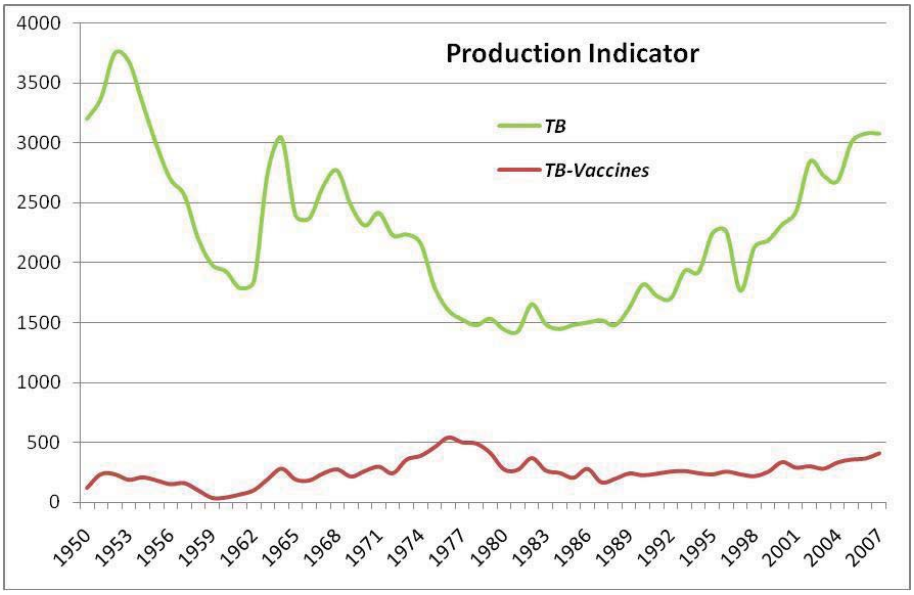


Figura 34: Evolución del indicador de producción en Tuberculosis ("*Tb*") y en Tuberculosis Vaccines ("*Tb Vaccines*"
)

de 1,844 artículos en el año 1962. Después, esta cifra se fue incrementando anualmente hasta alcanzar un valor de 3,040 artículos en el año 1964. Durante los siguientes diez años (1964-74) se manifiestan altibajos de menor escala alrededor de un valor promedio de 2,459 artículos; después este nivel de producción disminuye rápidamente a 1,797 en 1975; durante el período 1975-88 el ritmo de producción se estabiliza, con pequeñas fluctuaciones, alrededor de un valor promedio de 1,528 artículos. A partir de 1989 se inicia una tendencia de crecimiento sostenido (con excepción poco significativa en 1997), llegando a alcanzar en enero de 2008 un valor de 3,077 documentos. La extrapolación de esta tendencia predice que en los próximos años se recupere el más alto nivel de producción de todo el período de estudio, es decir, el que se tuvo al inicio del mismo.

Algunos especialistas del campo han investigado y ofrecido explicaciones de algunas de las variaciones que ha tenido el valor del indicador bibliométrico de producción. Por ejemplo, se ha considerado que la aparición de cepas del M. tuberculosis resistentes a los medicamentos tradicionales y la epidemia de HIV motivan nuevas investigaciones diagnósticas y terapéuticas así como la urgencia de mejorar la BCG y/o desarrollar nuevas vacunas Sierra [2006] Ly and McMurray [2008].

El hecho de que la producción en el tema de vacunas sea significativa-

mente menor que la producción total en tuberculosis obedece al hecho de que existe una gran variedad de aspectos (además de las vacunas), que abarca la investigación relacionada con esta enfermedad, sin embargo podría también ser indicativo de la dificultades intrínsecas asociadas, tanto a la investigación teórica fundamental, como a la invención de técnicas inherentes al desarrollo de medidas de prevención y control, alternativas a las tradicionales.

Para analizar la dinámica de la producción científica, relativa a diferentes temas de investigación relacionados con vacunas contra la tuberculosis consideramos la colección de registros bibliográficos (documentos semiestructurados) etiquetados con el término MeSH "Tuberculosis Vaccines". Esta colección de documentos D será tratada como una secuencia temporal $D = \{D_t\}_{t \in T}$, donde D_t son los documentos publicados en el año t y T es el periodo de análisis, que en este caso es $[1950, 2010]$.

6.2. Análisis temporal de ocurrencias de palabras

Mientras que las estadísticas de un periodo proveen una descripción sintética del desarrollo de un tema científico particular, las series de tiempo proveen entendimiento de los cambios en los productos y de su impacto en el tiempo. En informetía, el análisis de series de tiempo se ha enfocado en aspectos como: monitoreo científico-tecnológico, identificación de tendencias, etc.

En este capítulo se propone un método que usa las series temporales de ocurrencia de descriptores para el análisis de la evolución de un dominio de conocimiento específico. Este método aprovecha las capacidades de agrupamiento y visualización de la red neuronal SOM. Los resultados que se presentan en esta sección se obtienen de confrontar el problema de analizar patrones temporales obtenidos del cómputo de ocurrencias de descriptores MeSH.

En la sección 6.2.1 se muestra el efecto que tiene sobre la proyección inducida por el SOM, la naturaleza estadística de las series temporales obtenidas del análisis de ocurrencia de los descriptores MeSH. En la sección 6.2.2 se describe una transformación de las series temporales de manera que series temporales que corresponden a términos con alta frecuencia sean comparables con series temporales de términos con baja frecuencia. Por último, en la sección 6.3 se propone una modificación del algoritmo original que permite obtener mejores resultados.

6.2.1. Efecto de la ley de potencia en el mapeo de secuencias temporales

En esta sección se muestran los resultados de una aplicación de un algoritmo SOM básico con métrica euclideana. Sean $\{D_t\}_{t\in T}$ un panel informétrico con $T = \{t_1, ..., t_n\}$ conjunto de años consecutivos, $D = \cup_{t\in T} D_t$ y sea $kw \in KW(D)$. La secuencia temporal de ocurrencias para kw está dada por $\vec{kw} = (kw^{t_1}, ..., kw^{t_n})$, tal que para cada $t \in T$

$$\vec{kw}^{\,t} = \sharp\{d \in D \mid kw \in KW(D_t)\}$$

A continuación se muestra el mapa obtenido al procesar las secuencias temporales asociadas al dominio de investigación biomédica *Tb Vaccines*. En este caso el período de estudio $T = \{1950, ..., 2008\}$, el número total de descriptores $\sharp KW(D) = 15,939$. El conjunto de secuencias temporales que resulta,

$$K\vec{W} = \{\vec{kw}\}_{kw\in KW(D)},$$

se consideran como vectores en $\mathbb{R}^{\sharp T}$ y se entrena un modelo de SOM básico.

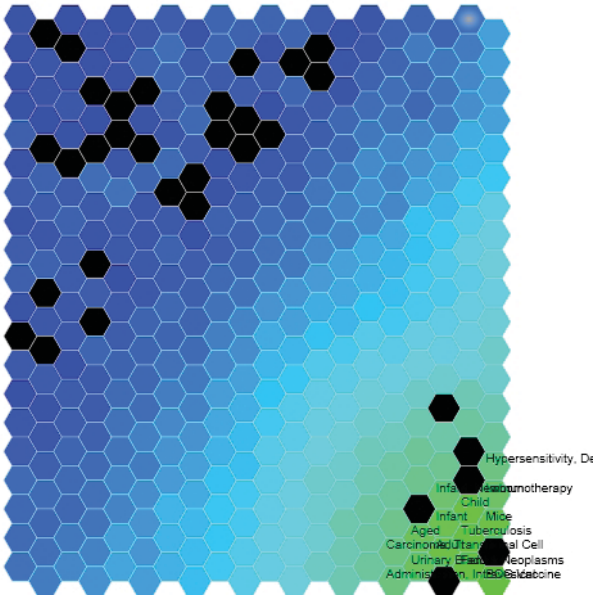


Figura 35: Resultado obtenido al entrenar un SOM básico con 400 neuronas utilizando las secuencias temporales de indicadores de actividad de los términos MeSH. La coloración corresponde al nivel de actividad total y las etiquetas son los términos MeSH más frecuentes.

La coloración de este mapa de la figura 35 lo determina la ocurrencia promedio de los descriptores que son proyectados en cada neurona. De manera que para el método de visualización se considera una barra cromática $mathcalC$ y queda definido de la siguiente manera:

$$\zeta : [0, \text{máx}\{\sum_{t \in T} \vec{kw}(t)\}_{kw \in KW(D)}] \rightarrow (C)$$

$$col(\eta) = \begin{cases} \zeta(\frac{\sum_{kw \in V_\eta} \sum_{t \in T} \vec{kw}(t)}{\sharp V_\eta}) & V_\eta \neq \emptyset \\ 0 & V_\eta = \emptyset \end{cases}$$

Aquellos descriptores que son proyectados en zonas verde corresponden a descriptores cuya ocurrencia es máxima, es decir, corresponde a descriptores principales. Esto se puede constatar al proyectar las etiquetas de los 30 descriptores con mayor frecuencia. Las zonas verde agua corresponde a las ocurrencias intermedias y las azules a las más bajas.

Del análisis visual del mapa en la figura 35 podemos concluir que:

- El mapa agrupa a los descriptores en función de su grado de ocurrencia.
- Este agrupamiento no refleja similitudes en las series temporales de indicadores de actividad.
- Los descriptores de mayor frecuencia son agrupados en la esquina inferior derecha.
- El resto de los descriptores se distribuye de manera tal que la ocurrencia total decrece en dirección de abajo hacia arriba y de izquierda a derecha.

De los resultados anteriores se concluye que la aplicación directa del algoritmo SOM Básico sobre el conjunto de series temporales de descriptores no aporta información relevante más allá de una agrupación en función de su ocurrencia total. Las causas de este comportamiento están siendo estudiadas, hasta el momento hemos identificado una ley de potencias (como la ley de Zift) ocasiona que la red neuronal proyecte los datos priorizando las diferencias en órdenes de magnitud sobre similaridades cualitativas entre los patrones de ocurrencia.

De acuerdo a lo establecido por la *Ley de Zift* es de esperarse que la distribución de los valores de las frecuencias cumplan una ley de potencia. Entre otras cosas esto implica que la mayoría de los términos MeSH ocurren con muy baja frecuencia y muy pocos son los que tienen las frecuencias

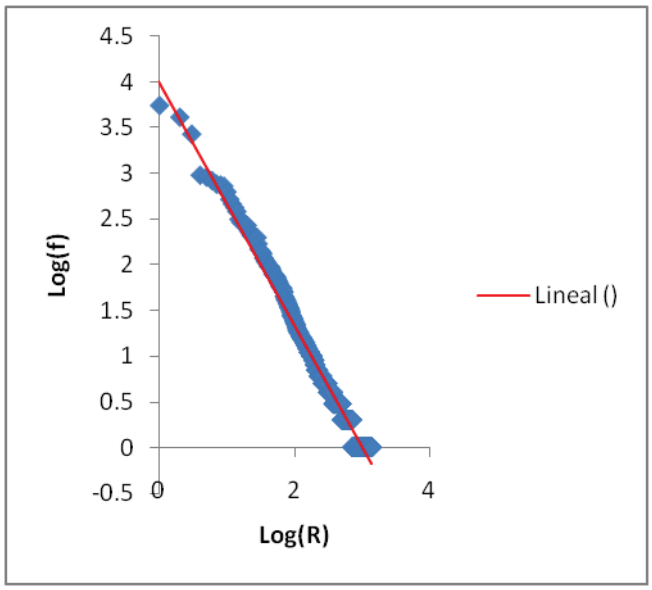


Figura 36: log – log gráfica de la distribución de la frecuencia de palabras clave en las vacunas contra la Tb. El eje horizontal correspondo con una al logaritmo del Rankeo (R) obtenido al ordenar las palabras de acuerdo al su frecuencia (f) y el eje vertical corresponde al logaritmo de f.

más altas Zipf [1935]. Como se puede ver en la figura 36, las frecuencias de los descriptores asociados al dominio *Tb Vaccines* siguen una distribución cercana a la ley de potencia.

Las concluciones del análisis e interpretación de estos resultados, suguieren que es necesario implementar modificaciones de la configuración básica del algoritmo SOM, para procesar series temporales tomando en cuenta las características propias del análisis informétrico.

6.2.2. Normalización de secuencias temporales

Como método para normalizar las secuencias temporales de ocurrencias seleccionamos la norma L_∞. Esta normalización será de utilidad para revelar e interpretar patrones encriptados en cursos temporales. De manera que para cada kw ahora se representa por un vector normalizado $\tilde{kw}$ de la forma:

$$\tilde{kw} = \left(\frac{1}{\text{máx}\left\{(\vec{kw}(t))\right\}_{t\in T}}\right)\vec{kw} \tag{6.3}$$

$$\tilde{KW} = \{\tilde{kw}\}_{kw\in KW(D)} \tag{6.4}$$

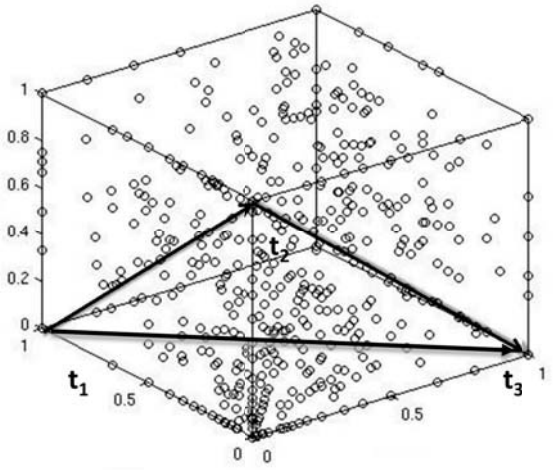


Figura 37: Proyección tridimensional del hipercubo obtenido de las secuencias temporales normalizadas.

Esta es una normalización no uniforme que define una proyección entre el espacio de entrada original de las secuencias temporales de enteros a la superficie de un hipercubo n-dimensional (ver la figura 37).

Una desventaja del uso de una norma L_n para la normalización, es la posibilidad de que, para periodos prolongados ($\#T \to \infty$), puede suceder que $L_n(kw) \to \infty$, esto implicaría que $kw/L_n(kw) \to 0$.

En periodos finitos, la normalización L_n tendría el efecto de aplastar la secuencia temporal, de tal manera que las palabras clave con mayor incidencia estarían más cerca de la secuencia nula, que aquéllas con incidencias más bajas, en tanto que en el caso de la normalización L_∞ la distancia geométrica mínima de la secuencia nula corresponde con aquéllos kw que sólo se presentan en un año y es distancia es de 1.

En la figura 38 se muestra el resultado de entrenar un SOM básico utilizando las secuencias temporales normalizadas. Este mapa utiliza el mismo código de color que el descrito para la figura 35.

Del análisis que se muestra en la figura 38 se puede apreciar que a pesar de que el efecto de la ley de Zift se atenúa, este se mantiene a pesar de la normalización. Este hecho motiva el replanteamiento del SOM básico, especialmente en lo referente a la manera en que se establece la similitud entre las secuencias temporales.

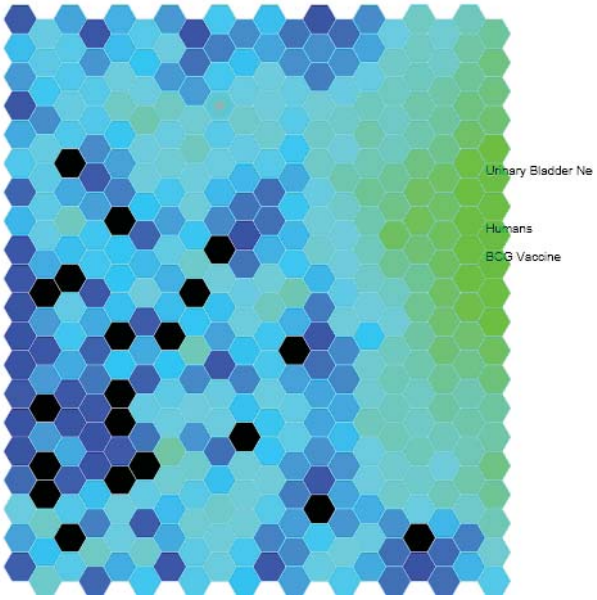


Figura 38: Resultado obtenido al entrenar un SOM básico con 400 neuronas utilizando las secuencias temporales normalizadas de indicadores de actividad de los términos MeSH.

6.2.3. Función de similitud para secuencias temporales de ocurrencias

La opción heurística para medir la similitud entre objetos que se representan como puntos espacios métricos multidimensionales están basadas en métricas euclidianas. Sin embargo, los resultados de las secciones anteriores indican que, para el caso de las secuencias temporales, el optar por una métrica euclidiana tiene como consecuencia que secuencias temporales que se encuentran cercanas en el hipercubo, no necesariamente representan descriptores que tuvieron una ocurrencia similar en el tiempo.

Por ejemplo, en la figura 37 los puntos $t_1 < t_2 < t_3$, que corresponden con los vértices del hipercubo, se encuentran a la misma distancia geométrica unos de otros. Por otra parte, estos puntos corresponden a años en la línea de tiempo, por lo tanto la distancia temporal entre t_1 y t_2 no es la misma que la que hay entre t_1 y t_3.

Para captar la naturaleza temporal de los datos de entrada es necesario incorporar un índice de disimilitud que no esté basado en una distancia geométrica. Si consideramos el dibujo de los patrones de la figura 39, la secuencia temporal a es cercana en el tiempo a b y de la misma manera que la secuencia c es cercana a b, supongamos que ambas distancias son iguales a d.

Partiendo de estas hipótesis podemos suponer que $d(a,b) = d(b,c) = \delta$. Si suponemos que la función de disimilaridad cumple la desigualdad del triángulo esto implicaría que $d(a,c) \leq d(a,b) + d(b,c) = 2\delta$; Sin embargo, si supone-

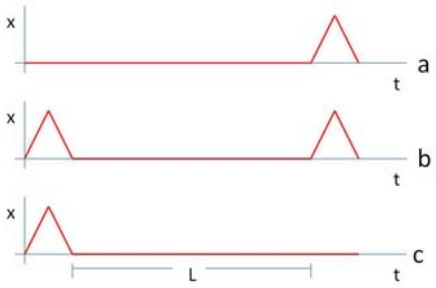


Figura 39: Comparación de patrones temporales

mos que $L \simeq d(a,c)$, dado que L es una distancia en el tiempo, podría ser lo suficientemente grande para que $d(a,c) > 2\delta$ y la desigualdad del triángulo no se cumpliría.

A diferencia de los criterios de competencia usados comúnmente, el criterio ad-hoc que se propone a continuación no resulta ser una función distancia (no cumple la desigualdad del triángulo). Diversas investigaciones han acumulado pruebas considerables de que los humanos y los animales dependen de landmarks (puntos clave) para organizar su memoria espacial Cheng and Spetch [1998]. Con base en esta noción en Parker [2000] se propone un modelo de similitud para las series de tiempo. Un landmark L se define como el punto de ajuste *(tiempos, eventos)* de mayor importancia.

Los autores consideran que estos puntos de mayor importancia son la máxima local y la mínima local y construyen una función de distancia bidimensional $\delta(L_1, L_2) = (\delta^{amp}, \delta^{temp})$, donde los landmarks se comparan en términos de distancia de amplitud y distancia en la línea de tiempo.

En este trabajo adoptamos una versión simplificada del enfoque de landmarks. Suponemos que todas las secuencias temporales normalizadas tienen el valor máximo de 1. La medida de disimilitud propuesta $\delta(\tilde{kw}^1, \tilde{kw}^2)$ para secuencias temporales normalizadas $\tilde{kw}^1, \tilde{kw}^2 \in \mathbb{H}$ se da por

$$\delta(\tilde{kw}^1, \tilde{kw}^2) = \begin{cases} d_\infty(\tilde{kw}^1, \tilde{kw}^2), & \arg\max\{\tilde{kw}^1\} \cap \\ & \arg\max\{\tilde{kw}^2\} \neq \emptyset \\ d_{set}(\arg\max\{\tilde{kw}^1\}, \arg\max\{\tilde{kw}^2\}), & \text{en otro caso} \end{cases} \tag{6.5}$$

donde $\arg\max\{\tilde{kw}\}$ es el conjunto de índices de tiempo donde x alcanza sus valores máximos, $d_\infty(\tilde{kw}^1, \tilde{kw}^2) = max_t\{|\tilde{kw}_t^1 - \tilde{kw}_t^2|\}$ es la distancia L_∞

y $d_H(\tilde{kw}^1, \tilde{kw}^2)$ es una función de distancia entre los conjuntos de puntos en un espacio normado (i.e. distancia de Hausdorff, enlace único o enlace completo). En este caso, el espacio normado es la línea de tiempo en la que se encuentran los puntos de los conjuntos $\arg\max\{\tilde{kw}\}$. A esta medida de similitud la denominaremos NMTDF (*Non-metric Temporal Dissimilarity Function*).

En la fórmula anterior (6.5) la distancia en amplitud se considera cuando las secuencias temporales comparten un punto de tiempo donde ambas llegan a su punto máximo y la distancia de tiempo se considera lo contrario. Para el último caso, seleccionamos la distancia de enlace simple dada por:

$$d_{set}(\arg\max\{\tilde{kw}^1\}, \arg\max\{\tilde{kw}^2\}) = |\min(\arg\max\{\tilde{kw}^1\}) - \min(\arg\max\{\tilde{kw}^2\})|. \tag{6.6}$$

En la siguiente proposición se resumen las propiedades básicas de la función de disimilitud propuesta. Como se verá, la opción de enlace único tiene una interpretación útil en términos de la gráfica de similitud inducida.

Proposición 2 *Supongamos $x, y \in \mathbb{H}$ y la medida de disimilitud (6.5) con función d_{set} dada por 6.6 entonces:*

1. *Si $\arg\max\{x\} \cap \arg\max\{y\} \neq \emptyset$, $0 \leq \delta(x, y) \leq 1$.*
2. *Si $\arg\max\{x\} \cap \arg\max\{y\} = \emptyset$, $1 \leq \delta(x, y) \leq |T|$.*
3. *La complejidad computacional para el cálculo de $\delta(x, y)$ es $O(\#T)$.*

6.3. Evolución en la investigación de Vacunas contra la Tuberculosis

En esta sección se muestra el mapeo cienciométrico que se obtiene como resultado de entrenar una red neuronal SOM, usando la medida de disimilaridad NMTDF definida en la ecuación (6.5). El resultado final es un mapeo en el cual los agrupamientos de descriptores se establecen de acuerdo a la similitud entre las secuencias temporales. Este mapeo resulta en una visualización que permite identificar distintos períodos en el desarrollo del campo de investigación. Estos períodos son caracterizados semánticamente utilizando el método propuesto en la sección 6.1.2.

En la sección 6.3.1 se muestran los agrupamientos obtenidos al aplicar el método de clustering *K-means* utilizando la NMTDF. Los agrupamientos obtenidos definen distintos momentos de la investigación relacionada con vacunas contra la tuberculosis. En la sección 6.3.2 se muestra el mapeo de los

descriptores MeSH asociados a la investigación en vacunas contra la tuberculosis, utilizando un SOM equipado con la NMTDF. La interpretación de los mapas obtenidos coincide con el análisis longitudinal que presentamos en el libro sobre vacunas contra la tuberculosís Guzmán et al. [2010]. En la 6.3.3 se presenta un mapeo de las principales líneas de investigación identificadas.

6.3.1. Obtención de agrupamientos y patrones temporales

A continuación se discute el resultado de procesar el conjunto de series temporales normalizadas $K\tilde{W}$ como en 6.4. El conjunto $K\tilde{W}$ es analizado aplicando un método de clustering basado en el método *K-means* configurado con la medida de similitud NMTSF.

En la obtención de un agrupamiento óptimo, para cada $k \in \{2, ..., 100\}$ se realizaron 10000 ejecuciones del algoritmo *K-means*, variando la condición inicial de los centroides. Se aplicó el criterio de Dunn para determinar el agrupamiento óptimo. Mediante la aplicación de este método se determinan 3 clusters (o clases) de descriptores. Los elementos de cada clase son términos MeSH que ocurren en Tb-Vaccines y siguen alguno de los patrones que se muestra en la figura 40.

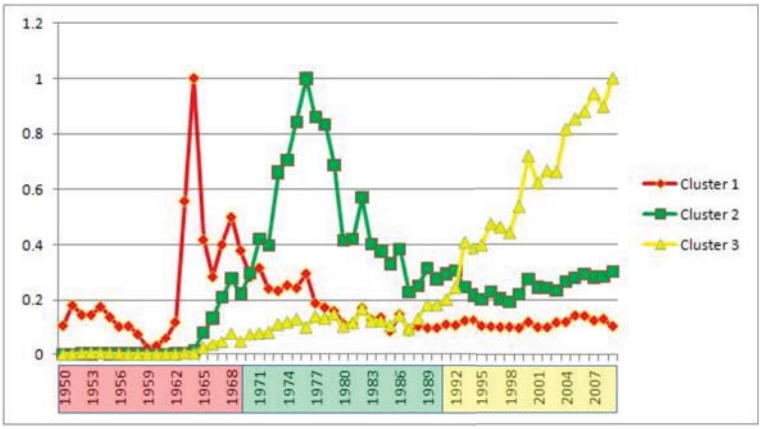


Figura 40: Patrones de los centroides obtenidos por el método K-means modificado con la medida de disimilaridad 6.5.

Como se puede apreciar en la figura 40, cada uno de estos patrones domina en algún intervalo de alrededor de 20 años. Las franjas de color sobre el eje de los años nos ayuda a identificar estos intervalos de tiempo y de esta manera asociar a cada cluster un subperíodo. Los descriptores en cada cluster se

caracterizan por presentar su principal actividad dentro de este subperíodo. De manera que se puede caracterizar semánticamente cada cluster (ver tabla en la figura 41) identificando los descriptores más ocurren.

Cluster	Periodo	Descriptores Principales	
1	1950-1968	*BCG Vaccine	*Research
		*Tuberculin Test	Mass Screening
		Tuberculosis: *prevention & control	*Tuberculin
		*Tuberculosis	Mass Chest X-Ray
		Tuberculosis: *immunology	Leprosy: *prevention & control
		BCG Vaccine: *complications	*Tuberculosis, Pulmonary
		*Mycobacterium bovis	*Tuberculosis Vaccines
		Isoniazid: therapeutic use	*Mycobacterium tuberculosis
		*Hypersensitivity, Delayed	BCG Vaccine: *history
		Sex Factors	*Communicable Disease Control
2	1970-1989	BCG Vaccine	*Immunotherapy
		BCG Vaccine: *therapeutic use	Clinical Trials as Topic
		Tuberculin Test	Combined Modality Therapy
		Time Factors	Mycobacterium bovis: *immunology
		BCG Vaccine: therapeutic use	Lymphocyte Activation
		Administration, Intravesical	Skin Tests
		BCG Vaccine: administration & dosage	Mycobacterium bovis: immunology
		Immunotherapy	Urinary Bladder Neoplasms: *therapy
		BCG Vaccine: immunology	Drug Therapy, Combination
		Immunity, Cellular	Immunization
3	1990-2009	BCG Vaccine: *administration & dosage	Prospective Studies
		BCG Vaccine: *adverse effects	Mycobacterium tuberculosis: *immunology
		Follow-Up Studies	*Vaccination
		Vaccination	Retrospective Studies
		BCG Vaccine: *immunology	Cells, Cultured
		Prognosis	Tuberculosis: prevention & control
		Risk Factors	Incidence
		Treatment Outcome	Disease Models, Animal
		Antitubercular Agents: therapeutic use	Mycobacterium tuberculosis: immunology
		Neoplasm Staging	Prevalence

Figura 41: Tabla de los 20 primeros descriptores más frecuentes en cada clase temporal

En la tabla de la figura 42 se muestran los resultados de rankear los descriptores en cada período de acuerdo al criterio GPR presentado en la sección 6.1.2. Como se pude apreciar la redundancia de los descriptores obtenidos por el método GPR es mucho menor de lo que se muestra en la tabla de la figura 41. Lo cual es un indicio de que esta manera de establecer la relevancia de los descriptores es más adecuada para caracterizar los distintos períodos identificados por el análisis temporal.

P1	GPR1	P2	GPR2	P3	GPR3
BCG Vaccine	0.251960	BCG Vaccine	0.118835	Urinary Bladder Neoplasms	0.001270
Child	0.058122	Tuberculosis	0.031062	Immunization Programs	0.001199
Infant	0.056443	Immunotherapy	0.019596	Tuberculin Test	0.000236
Tuberculosis	0.051408	Tuberculosis, Pulmonary	0.018983	T-Lymphocytes	0.000214
Tuberculin Test	0.025327	Mycobacterium bovis	0.017471	Tuberculin	0.000144
		Neoplasms	0.014426	Mycobacterium bovis	0.000140
		Tuberculin Test	0.012807	Cytokines	0.000128
		Macrophages	0.012041	Adjuvants, Immunologic	0.000125
		Urinary Bladder Neoplasms	0.011474	Leprosy	0.000108
		Melanoma	0.010808	Macrophages	0.000107

Figura 42: Descriptores más relevantes de acuerdo al método GPR

6.3.2. Visualización de Clustering Temporal

Una característica particular de los mapas obtenidos es que son susceptibles de una interpretación temporal. Dada la representación de descriptores como series temporales y su posterior normalización, los años, como etiquetas temporales, también pueden ser representados de esta manera. En este caso, la secuencia temporal que se obtiene es trivial, ya que tiene ceros en todas las entradas salvo en la que corresponde al año en cuestión. Utilizando esta manera de representar a los años se proyectan las etiquetas temporales en la red neuronal SOM previamente entrenada utilizando la medida de similitud no métrica (ver etiquetas temporales en azul de las figuras 43,44).

Como se puede apreciar en la visualización U-matrix que se muestra en la figura 43, los términos MeSH se agrupan en la red neuronal definiendo cúmulos de diversos tamaños (manchas naranja obscuro) separadas por contornos de neuronas borde (manchas amarillas). Estos cúmulos corresponden a conjuntos de secuencias temporales muy parecidas entre ellas. De hecho alrededor de cada etiqueta temporal se observa una de estas manchas, las agrupaciones al rededor de las etiquetas temporales corresponde a agrupaciones de aquellos descriptores MeSH que ocurrieron únicamente en ese año.

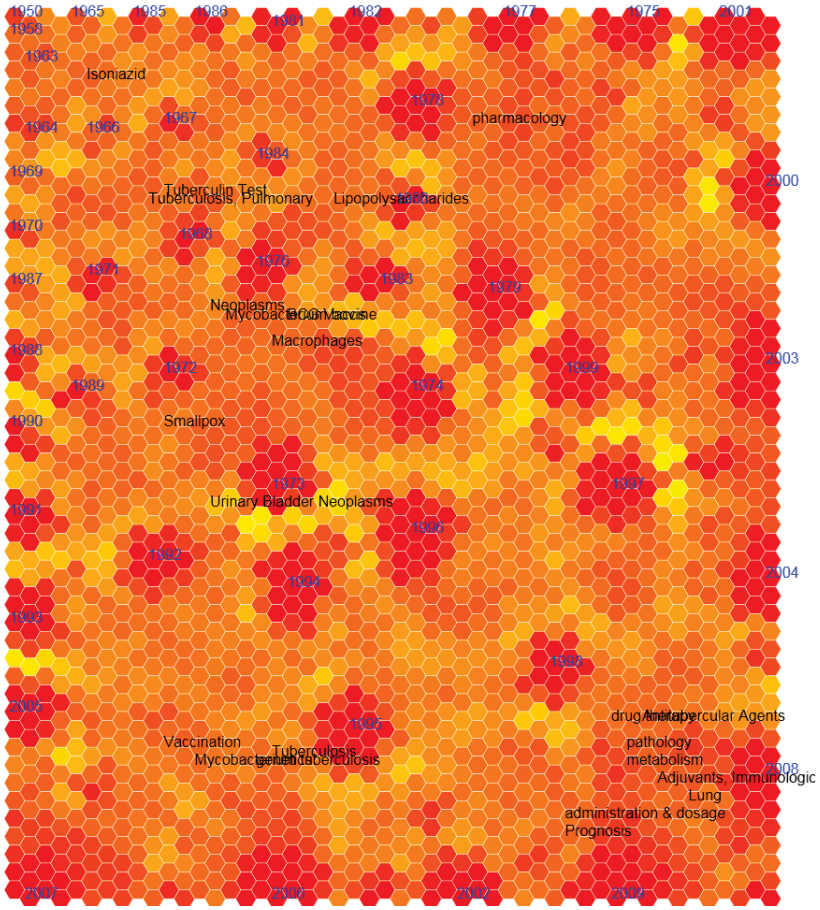


Figura 43: Visualización del mapeo temporal utilizando la técnica U-matrix

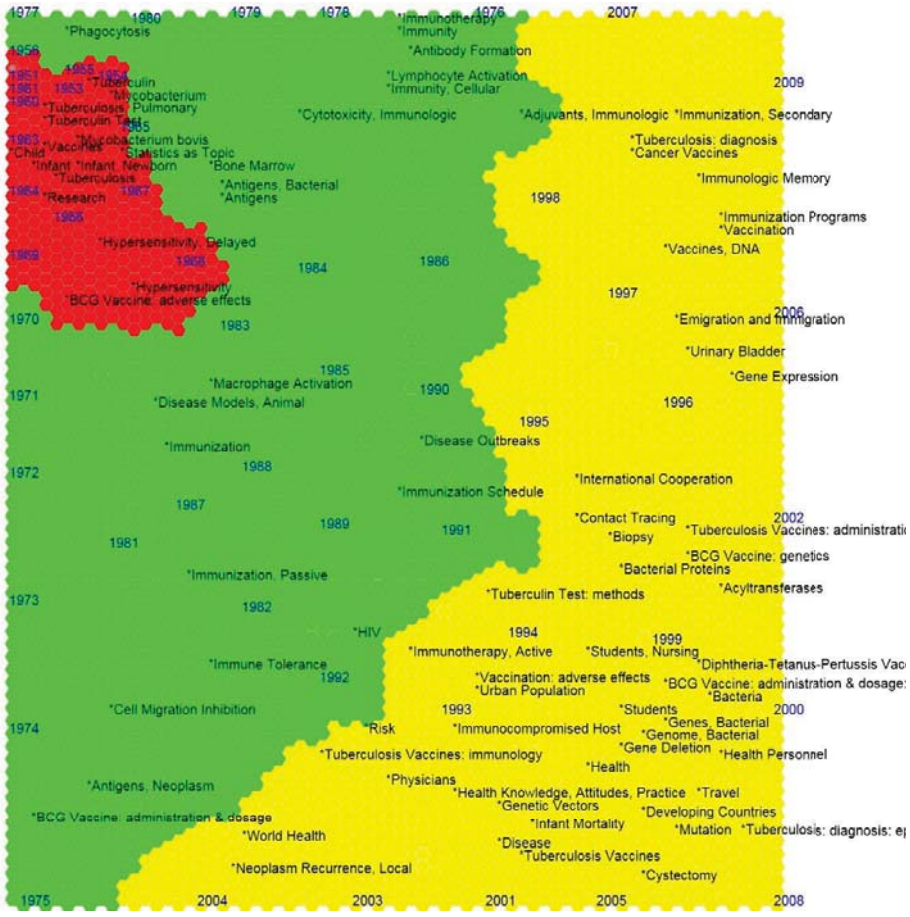


Figura 44: Proyección del clustering k-means temporal. Se utiliza el mismo código de color de la figura 40 y se proyectan las etiquetas de la tabla en la figura 41.

La coloración que se observa en el mapa de la Figura 44 está definida por las agrupaciones obtenidas en la sección 6.3.1, de manera que cada celda se colorea de acuerdo al cluster al que pertenecen los vectores de referencia. Como se puede apreciar las agrupaciones definen territorios conectados en el mapa de manera que la proyección preserva las relaciones de similitud establecidas por la agrupación.

Este hecho permite establecer relaciones de similitud entre los cursos temporales de los descriptores proyectados mediante su cercanía en el mapa. Si estos términos además de ocurrir en momentos similares, también se relacionan semánticamente se pueden establecer lineas de investigación que se siguen en el dominio de conocimiento. En la siguiente sección se muestra el

mapeo de las principales líneas de investigación identificadas en el dominio *Tb-Vaccines*.

En la siguiente subsección se realiza un análisis más detallado sobre estas relaciones de similaridad y la interpretación que estas relaciones tienen en términos de la evolución en la investigación en vacunas contra la tuberculosis.

6.3.3. Mapeo de las principales líneas de investigación

El mapa de la figura 45 pone de relieve el interés de la comunidad científica por temas específicos en distintos momentos. Estos temas definen líneas de investigación que evolucionan en el tiempo. En este mapa se identifican las principales lineas de investigación que se han trabajado a lo largo del periodo (1950-2009).

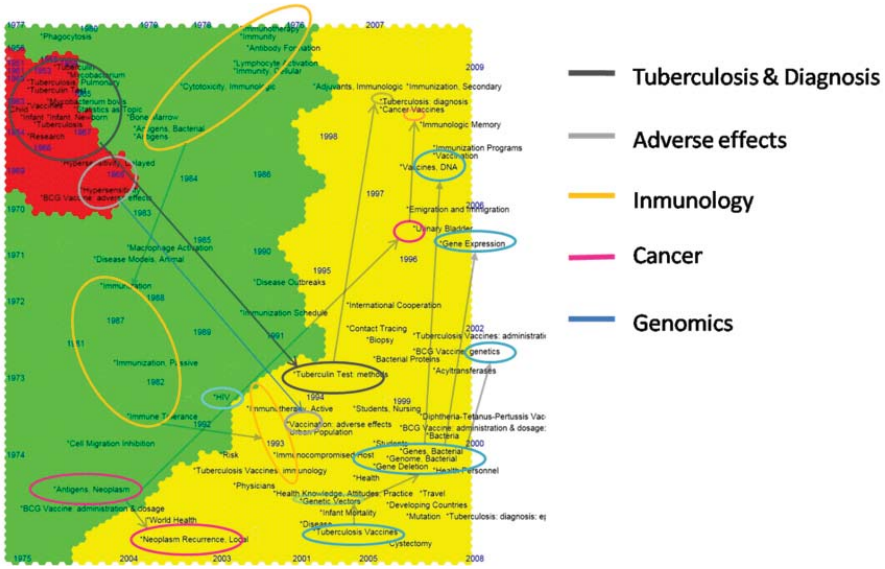


Figura 45: Mapa de la evolución en la investigación en Tb-Vaccines y principales frentes de investigación

Por otro lado, en la figura 46 se muestran los patrones temporales de las tres formas en las que aparece el término MeSH *BCG Vaccine*: ***BCG Vaccine** como término principal, **BCG Vaccine/*SubH OR SubH** cuando aparece con subheadings y **BCG Vaccine** cuando aparece sin marca adicional. Como se puede apreciar, cada forma en la que aparece este término tiene un patrón de occurrencia distinto. Esto es un indicio de la variación en el interés que la comunidad científica ha tenido entorno a este tema particular. Es de esperarse que el análisis de agrupamientos de secuencias temporales

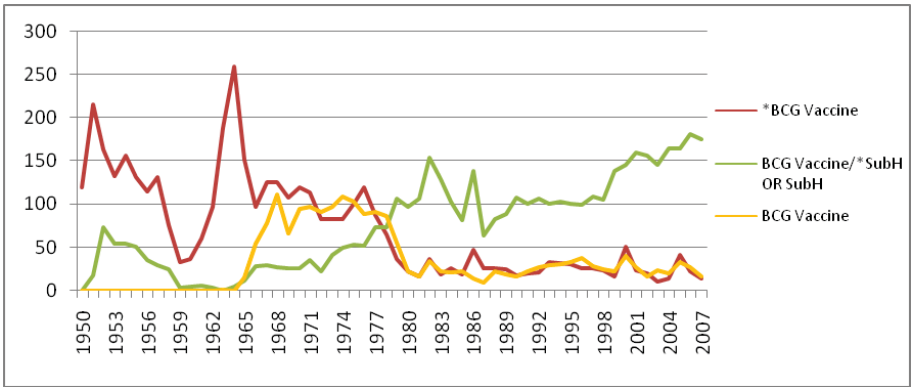


Figura 46: Desagregación de la producción relacionada con la vacuna BCG.

arroje patrones similares a las observadas en la figura 46 ya que como se ha mencionado la mayoría de la investigación del dominio *Tb Vaccines* está indexada con *BCG Vaccine*.

Estos resultados son útiles para establecer las distintas tendencias en la dinámica, ya que los descriptores se agrupan de acuerdo a patrón de ocurrencia en el tiempo. De esta manera es posible caracterizar distintos subperíodos por los temas principales que ocurrieron durante los mismos. También brinda la posibilidad de identificar aquellos descriptores temáticos que mantienen tendencias similares a los descriptores principales pero con un nivel de ocurrencia menor.

De la interpretación de este mapa se puede concluir que:

- las investigaciones asociadas con las líneas *Tuberculosis & Diagnosis* y *Adverse Effects* fueron las que concentraron el interés de la comunidad científica especializada en *Tb Vaccines*, durante el primer periodo identificado.
- la línea de investigación *Tuberculosis & Diagnosis* resurge en el tercer período.
- los temas relacionados con la línea *Inmunology* dominan el segundo período.
- un línea importante que emerge durante el segundo período es la de *Cáncer*
- el interés sobre *Cáncer* se mantiene durante el tercer período.
- es notable la emergencia del tema *HIV* justo en la frontera del segundo y tercer período.

- durante el tercer período es el claro interés sobre la línea *Genomics*, como estratégias para el desarrollo de una nueva vacuna.

6.4. Conclusiones desde la perspectiva de la KDViz

En este capítulo se presentan aportaciones relacionadas con la visualización de dominios de conocimiento en particular. Estas aportaciones se centran en identificar períodos que sirven para caracterizar distintos momentos en el desarrollo de la investigación y en analizar cómo evoluciona el interés de los investigadores en el tiempo.

El primer hallazgo presentado es que el uso de medidas de relevancia que consideran el conocimiento experto de los indizadores, es mucho más adecuado para caracterizar dominios de conocimiento ya que aportan una mayor ponderación a aquellos temas que son relevantes al tema en particular. Una prática común es considerar los descriptores de mayor frecuencia, este criterio suele ponderar descriptores muy generales que no describen el interés de científicos en un dominio específico. El método planteado pondera de manera adecuada los descriptores de manera que revela aquellos descriptores que aunque no son los más frecuentes, si son los más relevantes para el dominio de conocimiento en cuestión.

El segundo resultado relevante es que existe una dificultad intrínseca en procesar secuencias temporales de las frencuencias de descriptores derivado a la naturaleza estadística de estas frecuencias descritas por la *Ley de Zipf*. Esta característica estadística impone que el ordenamiento de los descriptores esté gobernado por su frecuencia y no por la similitud en su patrón de ocurrencia. Este efecto no se elimina con la normalización de las secuencias temporales.

Motivados por el efecto de la *Ley de Zipf* en la proyección y por la discusión presentada sobre lo poco adecuado de las funciones de distancia para representar las relaciones de similitud entre las secuencias temporales, presentamos una medida de similitud temporal no-métrica (NMTDF) la cual está inspirada en el concepto de *Landmarks*. Se muestra que esta función es útil para identificar patrones temporales que determinan distintos momentos en la evolución de la investigación en el dominio *Tb Vaccines*.

Los agrupamientos temporales obtenidos son caracterizados por los descriptores MeSH que resultan más relevantes de acuerdo al criterio propuesto. En la proyección de los descriptores es posible identificar subagrupamientos, los cuales coinciden en líneas de investigación previamente identificadas. Las líneas de investigación identificadas son consistentes con lo discutido con

expertos en el dominio de aplicación y reportado en Guzmán et al. [2010].

Capítulo 7

Conclusiones y trabajo futuro

En este trabajo se presentaron diversas aportaciones metodológicas en el campo de visualización y análisis informétrico. El principal rasgo distintivo de estas aportaciones es que todas ocupan como método analítico miembros de la familia SOM.

Las aportaciones metodológicas parten de una modelación matemática de las unidades analíticas (documentos semi-estructurados) y se concentran en una adecuada configuración de la red neuronal para llevar a cabo la minería de datos informétrica, así como un métodos de visualización que tienen como resultado visualizaciones más intuitivas y útiles para develar y representar el conocimiento intrínseco en los datos.

La principal conclusión de este trabajo es que la aplicación de modelos SOM para el análisis y visualización informétrica, requiere de consideraciones particulares para cada problema y cuanto más se considere la naturaleza de los datos a analizar en la cofiguración de los métodos, mejor será el resultado obtenido.

En el caso de estudio presentado en el capítulo 4 se establece una metodología para caracterizar patrones de desempeño de unidades productivas. Esta metodología tiene la ventaja de caracterizar patrones de desempeño e identificar datos extremos multivariados. Las visualizaciones provistas por la red neuronal permiten un análisis muy intuitivo de los distintos patrones de desempeño identificados y permite realizar una caracterización multiparamétrica de los mismos. Originalmente, los resultados de este método se presentan en Villaseñor et al. [2017] para la evaluación de la producción científica de Instituciones de Educación Superior. También se demostró su utilidad para la evaluación de revistas en Arencibia et al. [2016]. Como trabajo futuro se planea realizar otras aplicaciones del método en otros dominios temáticos del campo y de las ciencias sociales en general.

En el capítulo 5 se propone un método para visualizar nubes de puntos

híbridos. Este método se basa en la definición de una métrica pesada que determina una variable categórica control, variables numéricas dependientes y variables categóricas independientes. Los resultados obtenidos se basan en la observación de que, para el caso de variables categóricas binarias con distribución uniforme, el mapeo inducido por el SOM tiende a presentar una distribución simétrica. Esta propiedad es de gran utilidad para analizar poblaciones en las cuales se busque identificar asimetrías, por ejemplo el caso de los estudios con enfoque género como el que se reporta en Millán et al. [2012]. Como trabajo futuro se plantea la realización de otros ejercicios analíticos en donde se busque evidenciar las diferencias de género.

Por último, en el capítulo 6 se confronta el complejo problema de visualizar dominios de conocimiento. El desarrollo de los métodos propuestos fué motivado por el trabajo presentado en Guzmán et al. [2010]. Se muestra la necesidad de que las series temporales de ocurrencias de palabras, no sean tratadas como vectores y que la medida de similaridad implementada en el algoritmo SOM, no sea una función de distancia (métrica). La originalidad de la propuesta radica en el diseño de una medida de disimilitud temporal apropiada. Mediante el uso de la distancia temporal propuesta se obtienen mapeos que consideran la naturaleza temporal de los datos y permiten identificar distintos momentos en la evolución de dominios de conocimiento. El análisis de la evolucuón de dominios de conocimiento es una área de investigación de gran actualidad, a partir de lo que en esta tesis se ha presentado, se abre la posibilidad de aprovechar las capacidades de análisis y visualización de las redes SOM en esta tarea.

Bibliografía

R Arencibia, EA Villaseñor, I Lozano-Díaz, and H Carrillo. Elsevier's journal metrics for the identification of a mainstream journals core: A case study on mexico. *LIBRES: Library and Information Science Research Electronic Journal*, (1):1, 2016.

M. Berthold and D.J. Hand, editors. *Intelligent Data Analysis, an introduction*. Springer-Verlag, 1999.

J. Bertin. Graphics and graphic information processing. In Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman, editors, *Readings in information visualization*, pages 62–65. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

JP Bigus. *Data Mining with neural networks*. McGraw-Hill, New York, 1996.

Katy Börner, Chaomei Chen, and Kevin W Boyack. Visualizing knowledge domains. *Annual review of information science and technology*, 37(1):179–255, 2003.

Vicente P Guerrero Bote, Félix de Moya Anegón, and Victor Herrero Solana. Document organization using kohonen's algorithm. *Information processing & management*, 38(1):79–89, 2002.

R. J. Brachman and T. Anand. The process of knowledge discovery in databases. In *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA, USA, 1996. American Association for Artificial Intelligence. ISBN 0-262-56097-6.

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.

H Carrillo, Elio Villaseñor, and José-Luís Jimenez. Labsom, registro no. 03-2011-012713285100-01 ante el instituto nacional del derecho de autor, 2011.

J. Chappell and G. Taylor. The temporal kohonen map. *Neural Networks*, 6 (3):509–512, 1993.

Baitong Chen, Satoshi Tsutsui, Ying Ding, and Feicheng Ma. Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Journal of Informetrics*, 11(4):1175 – 1189, 2017. ISSN 1751-1577. doi: https://doi.org/10.1016/j.joi.2017.10.003. URL http://www.sciencedirect.com/science/article/pii/S1751157717300536.

C. Chen. *Information Visualization: Beyond the Horizon*, chapter Knowledge Domain Visualization. Springer-Verlag, London, 2006a.

Chaomei Chen. Top 10 unsolved information visualization problems. *IEEE computer graphics and applications*, 25(4):12–6, 2005.

Chaomei Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the Association for Information Science and Technology*, 57(3):359–377, 2006b.

K. Cheng and M. L. Spetch. *Mechanisms of landmark use in mammals and birds. Spatial representation in animals.* Oxford University, 1998.

Yizong Cheng. Convergence and ordering of kohonen's batch map. *Neural Computation*, 9(8):1667–1676, 1997.

Renato Corrêa and T. Ludemir. A hybrid som-based document organization system. In *SBRN'06, Ninth Brazilian Symposium on Neural Networks*, 2006.

Renato Corrêa and Teresa Ludermir. A quickly trainable hybrid som-based document organization system. *Neurocomputing*, 71(16-18):3353–3359, October 2008. doi: 10.1016/j.neucom.2008.02.021.

Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD*, KDD '01, pages 269–274, New York, NY, USA, 2001. ACM. ISBN 1-58113-391-X.

Joseph C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Cybernetics*, 3:32–57, 1973.

Leo Egghe. Expansion of the field of informetrics: Origins and consequences. *Information Processing & Management*, 41(6):1311–1316, 2005.

Usama Fayyad, Gregory Piatetsky-shapiro, and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, pages 37–54, 1996.

Paulo Henrique Muniz Ferreira and Aluízio Fausto Ribeiro Araújo. Modular self-organizing control for linear and nonlinear systems. In *Advances in Self-Organizing Maps and Learning Vector Quantization*, pages 131–141. Springer, 2016.

M.V. Guzmán. *ViBlioSOM: Metodología para la Visualización de Información Métrica con Mapas Auto-organizados*. Cuba, 2010.

MV Guzmán, H Carrillo, JL Jiménez, and EA Villaseñor. *The Art and Science of Tuberculosis Vaccines.*, chapter 22: Bioinfometric studies in TB vaccines research. Oxford Press, 2010.

M. Hagenbuchner, A. Sperduti, and A. C. Tsoi. A self-organizing map for adaptive processing of structured data. *IEEE Trans. on Neural Networks*, 14(3):491–505, 2003.

J. L. Jiménez-Andrade, E. A. Villase nor García, and H. Carrillo N. Una herramienta computacional para el análisis de mapas autoorganizados. In *IEEE 5 Congreso Internacional en Innovación y Desarrollo Tecnológico, At Cuernavaca, Morelos, México*, 2007.

Daniel A Keim, Florian Mansmann, and Jim Thomas. Visual analytics: how much visualization and how much analytics? *ACM SIGKDD Explorations Newsletter*, 11(2):5–8, 2010.

T. Kohonen. Analysis of a simple self-organizing process. *Biological Cybernetics*, 44(2):631–635, July 1982a.

T. Kohonen. Self-organizad formation of topologically correct feature maps. *Biological Cybernetics*, 43, 1982b.

T. Kohonen. *Self-Organizing Maps*. Springer, Verlag, New York, 1995.

T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela. Self organization of a massive document collection. *Neural Networks, IEEE Transactions on*, 11(3):574–585, may 2000. ISSN 1045-9227. doi: 10.1109/72.846729.

T. Koskela, B. Otaniemi, V. Markus, J. Heikkonen, and K. Kaski. Time series prediction using recurrent som with local linear models. *Int. J. of Knowledge-Based Intelligent Engineering Systems*, 2(1):60–68, 1998.

B. Y-L Kuo, T. Hentrich, B. M. Good, and Mark D. Wilkinson. Tag clouds for summarizing web search results. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 1203–1204, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7.

A. Langville and C. Meyer. The use of linear algebra by web search engines. *Bulletin of the International Linear Algebra Society*, 55:2–6, 2004.

Min Lu, Siming Chen, Chufan Lai, Lijing Lin, and Xiaoru Yuan. Frontier of information visualization and visual analytics in 2016. *Journal of Visualization*, May 2017. ISSN 1875-8975. doi: 10.1007/s12650-017-0431-9. URL https://doi.org/10.1007/s12650-017-0431-9.

L.H. Ly and D.N. McMurray. Tuberculosis: vaccines in the pipeline. *Expert Rev Vaccines*, 7(5):635–650, 2008.

Qiaozhu Mei and C.X. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 198–207. ACM, 2005.

Valeria Millán, EA Villasenor, N Martinez de la Escalera, and H Carrillo. Gender differences in academic performance at u.n.a.m. *Resources for Feminist Research*, 34(1/2):149, 2012.

V Millán, EA Villaseñor, N Martínez de la Escalera, and H Carrillo. *Visualización infométrica de algunas diferencias de género impresas en el egreso universitario.*, chapter VII, pages 191–210. Universidad de Colima, 2008.

Janne Nikkilä, Petri Törönen, Samuel Kaski, Jarkko Venna, Eero Castrén, and Garry Wong. Analysis and visualization of gene expression data using self-organizing maps. *Neural networks*, 15(8):953–966, 2002.

Ed Noyons. Bibliometric mapping of science in a policy context. *Scientometrics*, 50:83–98, 2001. ISSN 0138-9130.

Madalina Olteanu and Nathalie Villa-Vialaneix. Sparse online self-organizing maps for large relational data. In *Advances in Self-Organizing Maps and Learning Vector Quantization*, pages 73–82. Springer, 2016.

Chang-Shing Perng and Haixun Wang and Sylvia R. Zhang and D. Stott Parker. Landmarks: A new model for similarity-based pattern querying in time series databases. In *Proceedings of the 16th International Conference on Data Engineering*, pages 33–. IEEE Computer Society, 2000. ISBN 0-7695-0506-6.

Jean-François Pessiot, Young-Min Kim, Massih R. Amini, and Patrick Gallinari. Improving document clustering in a learned concept space. *Information Processing and Management*, 46(2):180–192, 2010. ISSN 0306-4573. doi: DOI:10.1016/j.ipm.2009.09.007.

Do Phuc and Mai Xuan Hung. Using som based graph clustering for extracting main ideas from documents. In *Research, Innovation and Vision for the Future, 2008. RIVF 2008. IEEE International Conference on*, pages 209–214, july 2008. doi: 10.1109/RIVF.2008.4586357.

Jonas Poelmans, Paul Elzinga, Stijn Viaene, and Guido Dedene. Curbing domestic violence: instantiating c–k theory with formal concept analysis and emergent self-organizing maps. *Intelligent Systems in Accounting, Finance and Management*, 17(3-4):167–191, 2010.

Xavier Polanco, Claire François, and Jean-Charles Lamirel. Using artificial neural networks for mapping of scienceand technology: A multi-self-organizing-maps approach. *Scientometrics*, 51:267–292, 2001. ISSN 0138-9130. 10.1023/A:1010537316758.

Jose C Principe, Ludong Wang, and Mark A Motter. Local dynamic modeling with self-organizing maps and applications to nonlinear system identification and control. *Proceedings of the IEEE*, 86(11):2240–2258, 1998.

Daniel Pullwitt. Integrating contextual information to enhance som-based text document clustering. *Neural Netw.*, 15:1099–1106, October 2002. ISSN 0893-6080. doi: 10.1016/S0893-6080(02)00082-5.

V.G. Sierra. Is a new tuberculosis vaccine necessary and feasible? A Cuban opinion. *Tuberculosis*, 86((3-4)):169–78, 2006.

André Skupin. From metaphor to method: Cartographic perspectives on information visualization. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, pages 91–97. IEEE, 2000.

André Skupin and Sara Irina Fabrikant. Spatialization methods: a cartographic research agenda for non-geographic information visualization. *Cartography and Geographic Information Science*, 30(2):99–119, 2003.

André Skupin, Joseph R Biberstine, and Katy Börner. Visualizing the topical structure of the medical sciences: a self-organizing map approach. *PloS one*, 8(3):e58779, 2013.

Panu Somervuo and Teuvo Kohonen. Self-organizing maps and learning vector quantization for feature sequences. *Neural Processing Letters*, 10 (2):151–159, 1999. ISSN 1573-773X. doi: 10.1023/A:1018741720065. URL http://dx.doi.org/10.1023/A:1018741720065.

Gilberto Sotolongo, Sánchez, María Victoria Guzmán, and Humberto Carrillo. Vibliosom: Visualización de información bibliométrica mediante el mapeo autoorganizado. *Revista Española de Documentación Científica*, 25 (4):477–484, 2002.

Marc Strickert and Barbara Hammer. Merge som for temporal data. *Neurocomputing*, 64:39–71, 2005. ISSN 0925-2312. doi: DOI:10.1016/j.neucom.2004.11.014.

Jean Tague-Sutcliffe. An introduction to informetrics. *Information processing & management*, 28(1):1–3, 1992.

J. W. Tukey. *Exploratory Data Analisys*. Addison Wesley, 1977.

Alfred Ultsch and Lutz Herrmann. The architecture of emergent self-organizing maps to reduce projection errors. In *ESANN*, pages 1–6. Citeseer, 2005.

Alfred Ultsch and Jörn Lötsch. Machine-learned cluster identification in high-dimensional data. *Journal of Biomedical Informatics*, 66:95–104, 2017.

P. Vanco and I. Farkas. Experimental comparison of recursive self-organizing maps for processing tree-structured data. *Neurocomputing*, 73(7-9):1362 – 1375, 2010. ISSN 0925-2312. doi: DOI:10.1016/j.neucom.2009.12.004.

Jarkko Venna and Samuel Kaski. *Neighborhood Preservation in Nonlinear Projection Methods: An Experimental Study*, pages 485–491. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001. ISBN 978-3-540-44668-2. doi: 10.1007/3-540-44668-0_68. URL http://dx.doi.org/10.1007/3-540-44668-0_68.

J Vesanto. SOM-based visualization methods. *Intelligent data analysis*, 1999.

Juha Vesanto and Esa Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on neural networks*, 11(3):586–600, 2000.

EA Villaseñor, R Arencibia, and H Carrillo. Multiparametric characterization of scientometric performance profiles assisted by neural networks: A study of mexican higher education institutions. *Scientometrics*, 110(1):77–104,

January 2017. ISSN 0138-9130. doi: 10.1007/s11192-016-2166-0. URL https://doi.org/10.1007/s11192-016-2166-0.

E Villaseñor, MV Guzmán, and Carrillo H. *Análisis de la dinámica de la relevancia de términos MeSH y su aplicación para la visualización de traslaciones en dominios de conocimiento bio-médico*. 2012. ISBN 978-959-234-076-3.

EA Villaseñor, H Carrillo, N Martinez de la Escalera, and V Millán. The use of weighted metric som algorithm as a visualization tool for demographic studies. 40:13–25, 2008a.

EA Villaseñor, Carrillo H, N Martinez de la Escalera, and Cruz N. Sistemas dinámicos y visualización informétrica: Una aplicación de la red neuronal som. 2008b.

EA Villaseñor, MV Guzmán, and Carrillo H. *Aplicación de ViBlioSOM al comportamiento temporal de los MeSH de MedLine*. 2010. ISBN 978-959-234-076-3.

Th Villmann, R Der, and Th Martinetz. A novel approach to measure the topology preservation of feature maps. In *ICANN'94*, pages 298–301. Springer, 1994.

Thomas Voegtlin. Recursive self-organizing maps. *Neural Networks*, 15(8-9): 979–991, 2002. ISSN 0893-6080. doi: DOI:10.1016/S0893-6080(02)00072-2.

H.D. White, X. Lin, and J. Buzydlowski. Co-cited author maps as real-time interfaces for web-based document retrieval in the humanities. In *Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*, 2001.

Sitao Wu and Tommy WS Chow. Prsom: A new visualization method by hybridizing multidimensional scaling and self-organizing map. *IEEE Transactions on Neural Networks*, 16(6):1362–1380, 2005.

Yang Xu, Lu Xu, and Tommy WS Chow. Pposom: A new variant of polsom by using probabilistic assignment for multidimensional data visualization. *Neurocomputing*, 74(11):2018–2027, 2011.

G. K. Zipf. *The psycho-biology of language*. Houghton, Mifflin., Oxford, England, 1935.