

## EL BIG BANG DE LOS DATOS (Parte II)

**Por: Juan José Lloret y Alejo Pomares**

Licenciados en Ciencias Matemáticas en la Facultad de Ciencias Exactas y Naturales  
– Universidad de Buenos Aires – Argentina

**ALEPH ZERO** Marketing & Social Research - **LYNX** Estadísticas & Sistemas –

**w<sup>3</sup>.TextualData.info**

## **Indice**

[Abstract y Key words](#)

1. [Planteo de la ponencia](#)

2. [Una intervención multidisciplinaria](#)

3. [Desarrollo del método para un caso.](#)

3.1. [Etapa de Representación de los datos](#)

3.2. [El ciclo de la intervención](#)

3.3. [Reducción de datos](#)

4. [Clasificación y búsqueda de conocimiento](#)

5. [Conclusiones](#)

[Apéndice](#) The Self-Organizing Map (SOM) por Teuvo Kohonen

[BIBLIOGRAFIA](#)

## **EL BIG BANG DE LOS DATOS (Parte II)**

Cuando el investigador en cualquier disciplina indaga sobre un fenómeno real que desea analizar y modelar, debe dar cuenta de grandes volúmenes de datos de características muy heterogéneas. En efecto, su mirada se enfrenta con una masa de datos con muy diferente estructura: textos, números, relaciones, eventos, imágenes, sonidos, sensaciones, etc. Es un desafío para el investigador buscar las herramientas que le permitan extraer de esa abigarrada colección de datos algo que pueda ser identificado como conocimiento agregado. La investigación de mercado no escapa a este fenómeno por el caudal de datos muy diversos y los requerimientos de optimizar tiempos y costos, y maximizar utilidades. Esto nos demanda desarrollar técnicas cada vez más refinadas de preparación, análisis e interpretación de los datos.

La presente ponencia intenta mostrar una serie de avances que estamos realizando en la etapa posterior a la obtención de los datos. En dichos desarrollos se conjugan distintas disciplinas que tratan de aportar una visión más amplia al fenómeno de la explosión de la información. En este estadio del proyecto, el énfasis ha sido condensado en el manejo de datos textuales y en la extracción de información y conocimiento para este tipo de datos. El sistema desarrollado está conformado por distintas etapas: 1) Conversión de los datos a algún sistema de estructura que pueda ser tratada por técnicas de ordenamiento documental 2) Búsqueda de una representación de los datos en algún espacio "geométrico" que permita visualizarlos y a la vez ensayar reglas que establezcan relaciones entre los datos. 3) Detección de patrones y construcción de un mapeo de conceptos que refleje la "realidad" al menos de una manera plausible y asimilable por el investigador.

**Palabras claves:** big bang de datos, corpus textuales, paradigmas, patrones, experto del dominio, red adaptativa, red semántica, red neuronal, clasificadores, datos multi-fuentes

When researchers in any discipline probe into a real phenomenon intended to be analyzed and modeled, they must account for large volumes of data with extremely heterogeneous characteristics. As a matter of fact, they are confronted with a myriad of data of very different nature: texts, numbers, relations, events, images, sounds, feelings, etc. Researchers, therefore, face the challenge to seek those tools that would enable them to draw from this entangled collection of data something that may be identified as value added knowledge. Market research is no exception to this phenomenon on account of the volume of very diverse data and the requirements for optimizing time and cost as well as maximizing profits. This situation calls for the development of increasingly refined techniques for data processing, analysis and interpretation.

This paper intends to show a number of advancements that are being made in the stage following data collection. Several disciplines converge into this process trying to provide a broadened outlook to the information explosion phenomenon. In this stage of the project, emphasis has been placed on the management of textual data and on data mining as well as knowledge acquisition for this type of data. The developed system is made up by different stages: 1) Data conversion into some structure system that may be managed by documentary classification techniques 2) Search for data representation onto some "geometrical" space that allows for their visualization while testing standards that set up relations among them. 3) Detection of patterns and construction of a concept mapping that reflects "reality" in such a way that can be at least plausible and apprehensible for researchers.

**Key words:** big bang of data, corpus text, paradigms, pattern, expert of domain, adaptive network, neural network, semantic network, classificatory methods, multi sources data.

## 1. Planteo de la ponencia

En sus primeros orígenes, la comunicación entre los anunciantes y los consumidores, era un proceso circular, donde el anunciante enviaba mensajes a los consumidores y recibía el feedback a través de la investigación clásica. En gran cantidad de veces, aun se usa un cuestionario, como un fenómeno estímulo/ respuesta y con un feedback desde el consumidor con forma estructurada y acotada.

Esta forma de monitorear al consumidor fue efectiva hasta la aparición de los *Internet juggernauts*<sup>1</sup>. A partir de allí el consumidor no solo pudo “expresar sus pensamientos” sino que dispuso de herramientas ágiles y efectivas para dar visibilidad a los mismos, a través de la ubicación en sitios, foros, diarios, blogs, etc. Esta tendencia posiblemente sea irreversible y global.

La avalancha de mensajes desde el anunciante, combinada con una caída de la confianza en ciertos paradigmas de comunicación y una baja en la colaboración por parte del consumidor, ha dado lugar al nacimiento de incipientes redes informales de comunicación que son necesarias estudiar.

Estas redes, verdadero entramado carretero donde circula la información entre las personas, tiene cierta analogía con las redes biológicas, por lo cual la comprensión en este ámbito del marketing de los fenómenos de propagación, estabilidad, memoria, excitación, etc. resultan de indiscutible utilidad no solo desde un punto de vista académico, sino también desde una óptica de preservación del negocio para la industria.

Muchos de estos aspectos de estudio, escapan al presente trabajo. En nuestro caso, nos hemos centrado en el análisis de ciertas características de los “objetos” que circulan en este tipo de red.

**Este trabajo ensaya un método de tratamiento de grandes corpus textuales,** situación típica en una comunicación escrita, que a su vez parte forma de los “corpus multimedia”<sup>2</sup>.

El esquema de desarrollo de este método se muestra a continuación.

---

<sup>1</sup> El término **juggernaut** es usado generalmente para describir cualquier fuerza literal o metafórica considerada como imparable y que avanza sobre todo en su camino.

<sup>2</sup> Usado como abuso de lenguaje

## RACIONAL DEL METODO



Como se observa en el diagrama el proceso de captación obedece a dos posibles modalidades: datos que provengan de un diseño de recolección o bien datos extraídos del flujo de la comunicación en la red y cuya composición es totalmente heterogénea. El campo con cuestionario es un ejemplo para el primer caso, mientras que en cambio la utilización de motores de búsqueda, alertas o robot proveerán datos de muy diferente organización.

Si proseguimos con el diagrama, la siguiente etapa está enfocada a obtener una adecuada Representación de la masa textual que se desea explorar. En esta etapa, el esfuerzo se concentra en hallar unidades de análisis "óptimas". Es en esta etapa en donde el experto de dominio juega un rol muy importante en la construcción de las modalidades, mediante la semantización de la masa textual inicial en categorías conceptuales con mayor tenor de conocimiento. Es aquí donde los grupos interdisciplinarios y los recursos de IT producen sinergia interesante. En esta etapa se produce la reducción de los datos: de miles o millones de palabras a unos pocos centenares de conceptos claves. Esta etapa sirve para sistematizar la información, evitar que la investigación se hunda obsesivamente en los datos (overfitting) impidiendo la detección de ordenamientos más "blandos" (soft) pero con mayor carga de conocimiento inferencial y por último, pero no por eso menos importante, permitir la optimización del proceso desde el punto de vista de recursos.

En la última etapa del método, se busca ganar conocimiento a través de revelar algún tipo de estructura en los datos, así como detectar relaciones o asociaciones conceptuales. Es en esta etapa donde se aplica el concepto de *machine learning*<sup>3</sup>,

<sup>3</sup> Ver por ejemplo: [http://en.wikipedia.org/wiki/Machine\\_learning](http://en.wikipedia.org/wiki/Machine_learning)

entrenando y testeando a través de clasificadores posibles estructuras de los datos que permiten obtener conocimiento.

## **2. Una intervención multidisciplinaria**

El presente trabajo se sitúa en un campo de intersección de varias disciplinas, algunas muy jóvenes y en formación, y otras que están en estadios más avanzados en su evolución: entre las últimas, la estadística, la psicología, la lingüística y el análisis del discurso, y entre las primeras, data y text mining, el procesamiento del lenguaje natural y la gestión de conocimiento. El enfoque multidisciplinario deviene al utilizar técnicas de análisis habitualmente reservadas para la información en formato “duro” a corpus textuales específicos, poco estructurados desde el punto de vista formal, pero ricos en contenido para el especialista del dominio origen de dicho discurso.

Es interesante mencionar aquí la evolución producida en el campo del procesamiento de datos que ha llevado a incorporar, por analogía biológica, el modelo actual de red neuronal y otros enfoques alternativos de modelización.

El paradigma que subyace a la mayoría de los sistemas de procesamiento de información puede resumirse en una fórmula sencilla: lógica dicotómica booleana + el algoritmo-máquina de Turing. Es decir, los estados “0” y “1” para representar los datos y reglas almacenables para operarlos. Este paradigma de “computación algorítmica” se aplica a la resolución de un problema a través de un algoritmo que se codifica como programa ejecutable (software) en un computador, se almacena en un dispositivo de memoria y se ejecuta en forma secuencial en un procesador o en varios en paralelo, según la complejidad y la necesidad de potencia de cálculo. Todo el proceso es guiado y controlado por una lógica subyacente que hace las veces de “cerebro” organizador y que asegura que la tarea se hará de acuerdo a lo ya planificado.

El éxito de este paradigma en campos muy diversos llevó a pensar en la posibilidad cierta de emular en un computador el razonamiento inteligente del cerebro humano. En los años 50's aparecieron los primeros programas para jugar al ajedrez, nace la “inteligencia artificial”, y el hito siguiente es la introducción de los sistemas expertos y la ingeniería de conocimiento. El sistema experto es la transformación en software de las reglas de decisión que un experto en el dominio utiliza para hacer su trabajo. Un sistema experto puede determinar la prima de un seguro, evaluar el riesgo crediticio,

identificar una avería entre millones de componentes, ubicar un producto en un depósito, guiar el vuelo de un avión, etc.

No obstante, pese a que estos sistemas expertos resuelven en segundos problemas que le resultarían irresolubles al cerebro humano o que le llevarían años de trabajo ininterrumpido, un ser viviente realiza casi sin esfuerzo tareas para las que todavía no hay ningún sistema experto capaz de hacerlas. Basta pensar en un niño que responde a una pregunta “inventando” una frase que nunca oyó ni dijo previamente o un adolescente reconociendo en segundos un nuevo tema de su grupo preferido.

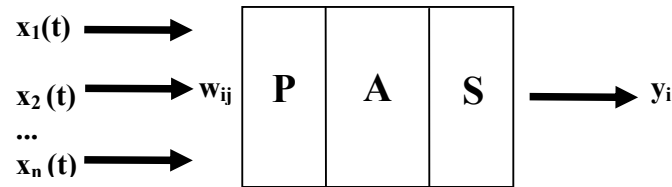
Hay múltiples evidencias que indican que el cerebro opera de manera diferente a los sistemas expertos de la ingeniería de conocimiento. Pese a que el mundo real le presenta la información de manera abigarrada y en grandes volúmenes, a veces muy imprecisa y distorsionada por “ruidos” diversos, un ser viviente puede en general y aún en las primeras etapas de su desarrollo biológico, resolver las dificultades que le plantea su entorno. Para ello parece disponer de un sistema de procesamiento con la capacidad de operar de manera distribuida y en paralelo, pero en un esquema adaptativo de cooperación y aprendizaje. El sistema “conoce y responde” a través de lo que produce la sinergia de miles o millones de nodos de una red operando de manera simultánea ante un estímulo, no de manera planificada y con control centralizado sino a través de la auto-organización y el entrenamiento a partir de su entorno próximo. Cada nodo realiza una tarea muy elemental que contribuye sinápticamente a la respuesta final.

Es por ello que, como alternativa al paradigma Boole -Turing, ha aparecido el enfoque de la computación evolutiva que reúne diversas técnicas que comparten el paradigma de replicar en modo artificial las soluciones que aparecen en el medio natural cuando los escenarios son “ruidosos”, con información en formatos muy diversos e imprecisos. Son ejemplos de estas técnicas los sistemas borrosos (fuzzy), las redes neuronales y los algoritmos genéticos.

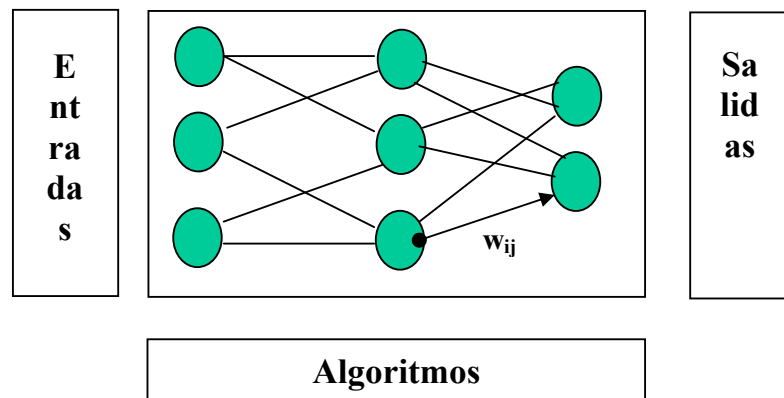
En el caso de las redes neuronales, y haciendo una analogía biológica (Del Brio y Molina), se denomina procesador elemental o neurona a un dispositivo simple de cálculo que a partir de un vector de entrada procedente del exterior o de otras neuronas, proporciona una única respuesta o salida.

En un instante  $t$ , los estímulos entrantes  $x_j(t)$  generan a través de los pesos  $w_{ij}$  la respuesta final  $y_i(t)$  de la neurona, y hay cierto consenso de los especialistas en distinguir tres procesos internos de:

- Propagación:  $p_i(t) = P(x_j(t), w_{ij})$
- Activación:  $a_i(t) = A(p_i(t))$
- Salida:  $y_i(t) = S(a_i(t))$



La red neuronal compuesta por estos procesadores elementales se representa matemáticamente como un grafo, consistente en un conjunto de nodos o vértices, y un conjunto de conexiones o “links” establecidas entre ellos. Las conexiones pueden indicar el sentido (flechas) en que se desplaza la actividad, y se suele indicar en cada conexión el peso pre-sináptico con que el nodo de salida estimulará al nodo de llegada. La aplicación de este modelo a un fenómeno en análisis puede resumirse en el siguiente:



La disponibilidad de herramientas más potentes de procesamiento se acompaña con un crecimiento vertiginoso de los corpus producidos y distribuidos por particulares y organizaciones vía internet o intranet corporativas. La necesidad de manipular y extraer conocimiento de esta masa creciente de datos hará frecuentes en el futuro las intervenciones como las aquí descritas, en busca de modelos de representación que le permitan al especialista del dominio en cuestión compactar y organizar la información disponible, catalogar los casos nuevos y predecir estados futuros del fenómeno en análisis.

La necesidad de manipular y extraer conocimiento de esta masa creciente de datos hará frecuentes en el futuro las intervenciones como las aquí descritas, en busca de modelos de representación que le permitan al especialista del dominio en cuestión compactar y organizar la información disponible, catalogar los casos nuevos y predecir estados futuros del fenómeno en análisis.

En los párrafos que siguen se desarrollará en detalle la aplicación de estos conceptos y técnicas a un corpus de datos textuales.

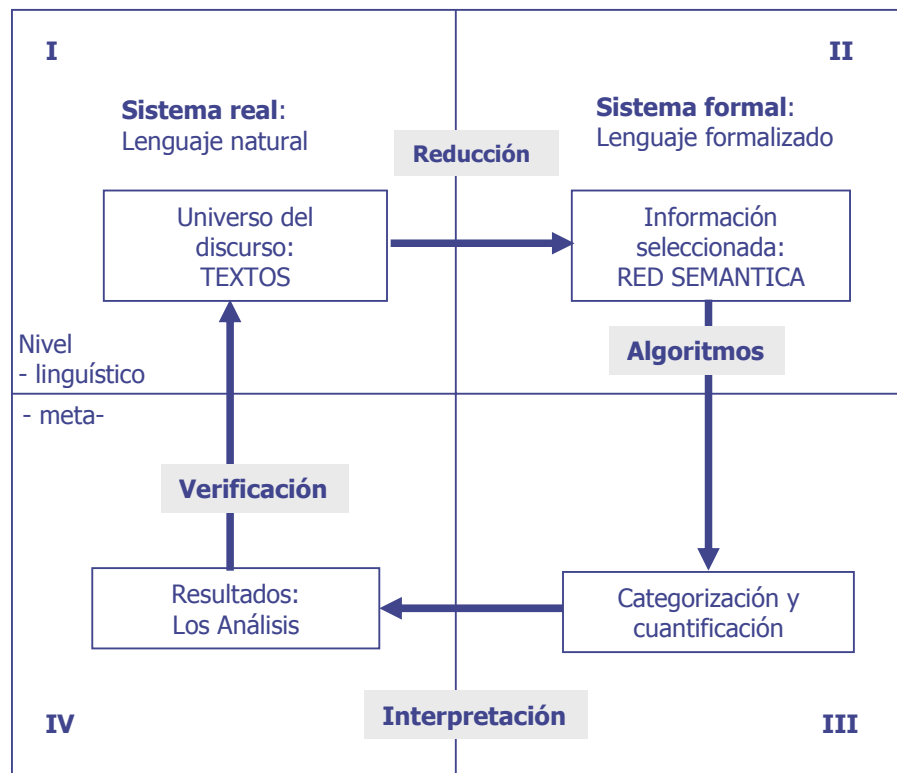
### **3. Desarrollo del método para un caso.**

#### **3.1. Etapa de Representación de los datos**

Para desarrollar el método propuesto, usamos un dataset construido con los verbatim recogidos durante varios años y de manera directa en las reuniones del Grupo Sostén para personas que han hecho al menos un intento suicida. Esta actividad de apoyo - no terapéutica - es coordinada por la Asociación Argentina de Prevención del Suicidio, y forma parte de su proyecto de "Redes Sociales de Apoyo en la Rehabilitación del Proyecto Suicida". Uno de los objetivos del proyecto es la reconstrucción de los lazos familiares y sociales del individuo a través del Grupo Sostén.

El material fue cedido por el Lic. Carlos Martínez, presidente de la Asociación, que además ha participado activamente en su carácter de experto en el dominio intervenido y a quien agradecemos la colaboración prestada a lo largo de todo el proyecto.

### 3.2. El ciclo de la intervención



Este gráfico resume el esquema general de una intervención sobre el corpus textual de un dominio.

Los cuadrantes superiores corresponden al lenguaje natural, en el primero en estado bruto ("raw-data") y en el segundo ya se ha operado una reducción, que en el caso de la intervención aquí descrita fue vía la lematización y la generación de una red semántica en colaboración con el experto en el dominio. Más adelante nos referiremos en detalle a estos procesos, pero es importante recalcar la función que cumplirá la red

semántica como guía de referencia en la extracción de contenidos y búsqueda de relaciones entre categorías conceptuales de mayor nivel de abstracción. Esta red o “escenario” es el tamiz a través del cual pasará el texto, revelando su entramado bajo el enfoque semántico del experto en el dominio.

El proceso de formalización – representación vectorial y algoritmo de transformación - produce un modelo formal en el dominio meta-lingüístico cuya interpretación nos devuelve al dominio real. La verificación del modelo determina su adecuación y su poder de representación de conceptualizaciones en el dominio en estudio.

### **3.3. Reducción de datos**

Es el proceso que extrae y pone en foco el vocabulario realmente presente en el corpus. Para este proceso se utilizó software desarrollado ad-hoc y Tropes Ver 6.0, un software específico para la indexación de textos y análisis cognitivo. Este paquete de análisis del discurso tuvo su origen en los desarrollos de la unidad de Análisis Proposicional del Discurso, Universidad París 8. (Ghiglione, Marchand y Otros).

El primer paso en la construcción de la red semántica es realizar el “parsing” del texto en párrafos, frases y lexicales elementales. El objetivo en esta etapa es presentarle al experto en el dominio un ranking con las frecuencias de aparición de cada lexical – en general a partir de un umbral mínimo de ocurrencia – con sus lemas asociados. Por un “lema” nos referiremos aquí a la palabra raíz de la que proviene el lexical, ej: dormí, dormía, dormiría, duerme, duermo, durmiendo → **dormir**

La tarea del experto será principalmente inductiva: en sucesivas etapas retro-alimentadas se irá generando la red semántica. Los procesos automáticos posteriores “aprenderán” de su *expertise* en el dominio en cuestión (paradigma ME = *machine learning*). Confróntese este enfoque, hoy muy frecuente en las aplicaciones, con el paradigma anterior de la “ingeniería de conocimiento”, en el cual el experto se concentraba en producir reglas de clasificación y análisis que luego se intentaba reproducir en los procesos automatizados (paradigma KE = *knowledge engineering*).

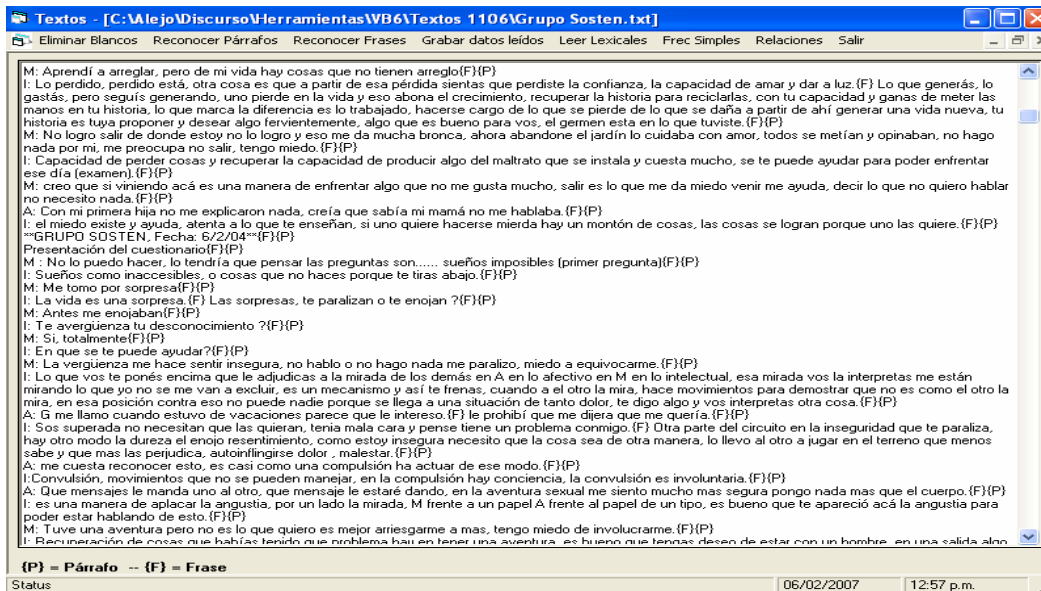


Figura 1. Parsing en párrafos y frases.

Para facilitar el tratamiento del texto, el software genera un informe en hipertexto que le permite al experto navegar el corpus, situándose en el contexto de cada aparición.

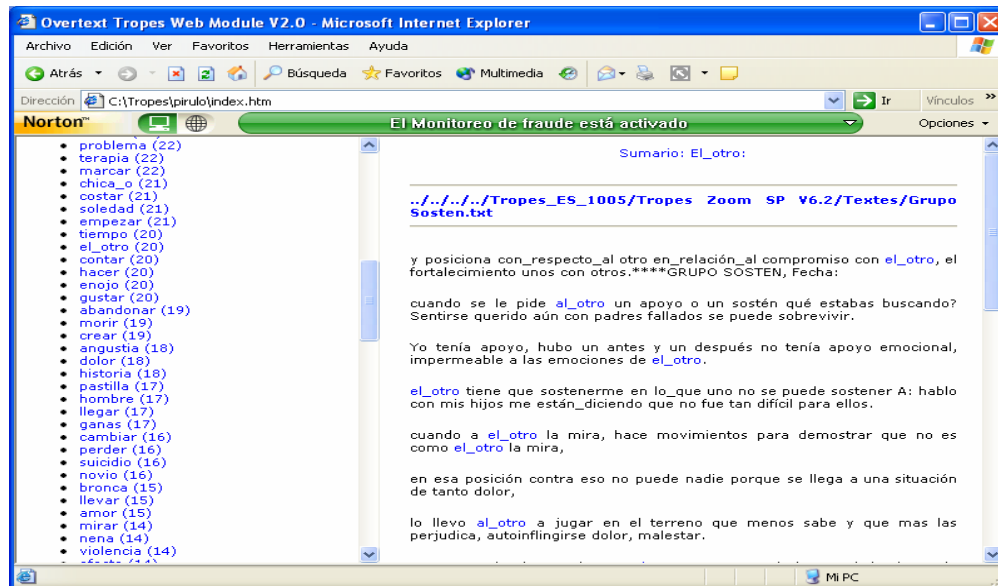


Figura 2. El corpus en hipertexto.

Durante el procesamiento se genera el análisis de puntuación y el sintáctico - gramatical, y la desambiguación lexical. Se estima que en los textos en idioma español hay en promedio más del 20% de referencias ambiguas, ya sean morfológicas (la palabra “vino” es una inflexión verbal y también una bebida) o semánticas ( el

sustantivo “aceite” se refiere a un aderezador, a un lubricante, a un producto cosmético, etc.). La calidad de la desambiguación dependerá de la riqueza de los diccionarios internos del proceso, y aquellas inherentes al dominio en cuestión pueden resolverse vía la red semántica que siempre tendrá prioridad en los análisis. El problema de la polisemia en la desambiguación es también un ejemplo de uso del paradigma ME: el proceso automático compara el contexto de ocurrencia de la palabra ambigua con contextos previamente desambiguados, y se decide por el de mayor similitud de uso. (Ej: “hice un cambio de aceite” produce, en general, la desambiguación del lexical “aceite”).

Este primer análisis ya es revelador de algunas características del texto tratado, que se hacen aún más evidentes al compararlo con un texto que hace uso de un lenguaje de contenido más neutro. Los “episodios” se generan con la introducción de un nuevo tema en el discurso o con la re-aparición de uno anterior, y los conceptos subyacentes a estos temas dan lugar a la aparición de nuevos nodos en la red semántica que caracterizan a las “frases relevantes”. La concentración de nodos en una frase y en su entorno próximo determina una mayor ponderación y por ende la hace más “relevante”. En este ejemplo el texto está confrontado con un corpus de noticias en las secciones el País, Opinión y Sociedad del diario Clarín.

CLARIN		GRUPO SOSTEN	
<b>Tamaño de la muestra en palabras:</b>			
515,402		19,532	
<b>Características del texto:</b>			
- estilo dominante: descriptivo		- estilo dominante: argumentativo	
- puesta en escena: dinámica, activa		- puesta en escena: anclada en lo real	
1423 frases relevantes	2.8	154 frases relevantes	<b>7.9</b>
89 episodio(s)	1.7	23 episodio(s)	<b>11.8</b>
<b>Verbos utilizados:</b>			
Factivo	65.60%	Factivo	43.59%
Estativo	27.00%	Estativo	<b>47.96%</b>
Declarativo	7.50%	Declarativo	8.76%
<b>Modalizaciones:</b>			
Tiempo	24.90%	Tiempo	16.79%
Lugar	6.50%	Lugar	6.47%
Modo	9.80%	Modo	4.08%
Afirmación	1.60%	Afirmación	0.40%
Duda	0.80%	Duda	0.96%
Negación	28.80%	Negación	<b>48.92%</b>
Intensidad	27.60%	Intensidad	22.38%
<b>Adjetivación :</b>			
Objetivo	45.80%	Objetivo	28.07%
Subjetivo	23.00%	Subjetivo	<b>39.14%</b>
Numérico	31.20%	Numérico	32.79%
<b>Pronombres :</b>			
Yo	6.80%	Yo	<b>62.04%</b>
Tú, vos, usted	1.00%	Tú, vos, usted	8.63%
Él, ella	3.10%	Él, ella	7.76%
Nosotros, nosotras	1.70%	Nosotros, nosotras	0.53%
Vosotros, ustedes	0.30%	Vosotros, ustedes	0.73%
Ellos, ellas	3.70%	Ellos, ellas	1.06%
Se	<b>79.50%</b>	se	19.24%

De la “densidad semántica” del texto del Grupo Sostén da cuenta la relación 2.8 vs. 7.9 en la tasa que describe la distribución de frases relevantes cada 1.000 palabras, y 1.7 vs. 11.8 en la correspondiente a episodios cada 10.000. Es un texto producido desde la subjetividad y lo emocional, el “ser” y el “estar”.

Las modalizaciones (adverbios o locuciones adverbiales) le permiten al locutor implicarse en lo que dice, situarlo en el tiempo y en el espacio, o darle intensidad y sentido. En el caso del uso de las diversas formas de la negación – no como mecanismo psíquico sino como forma de expresión – es relevante su persistencia a lo largo de todo el corpus analizado.

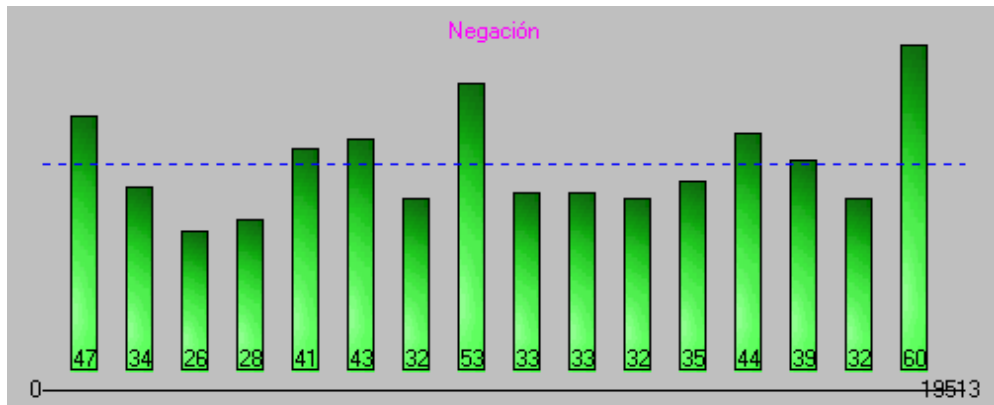


Figura 3. La Negación: frecuencias en una partición del corpus.

Con cada modificación introducida en la red semántica de soporte, se reprocessa el texto y todos los indicadores asociados para que el experto evalúe el impacto producido.

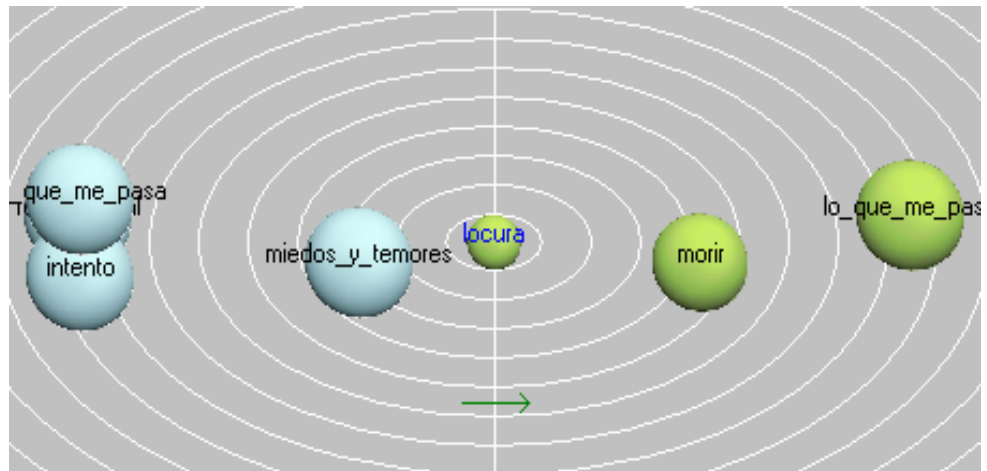


Figura 4. Una referencia, relaciones e intensidades.

El tamaño de la esfera graficada es proporcional a la frecuencia de aparición de la categoría en el discurso, y la distancia entre “órbitas” es un indicador de la intensidad de la relación.

En sucesivos reprocessos el experto va generando la red semántica que dará soporte al modelo de representación conceptual.

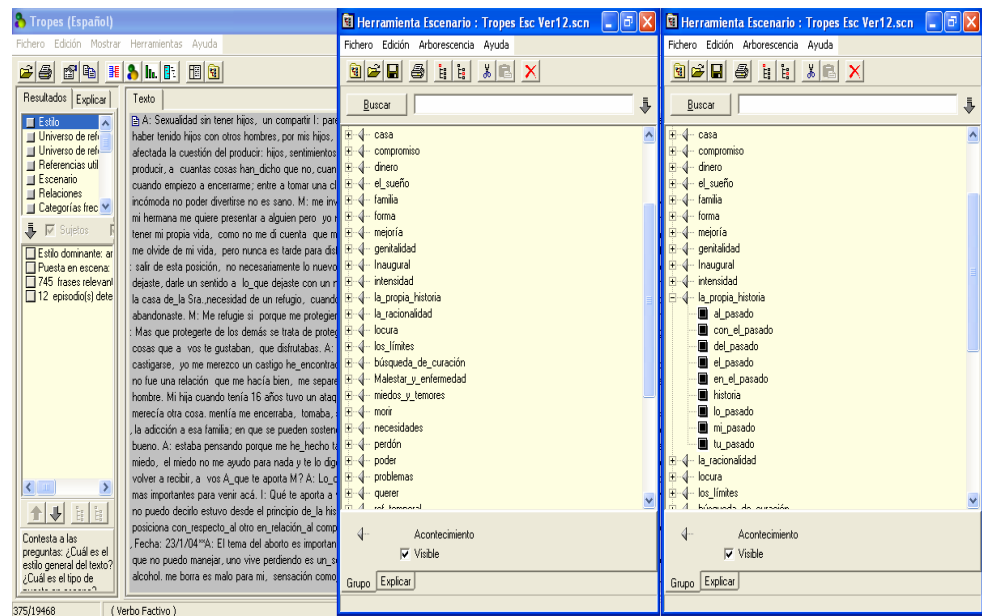


Figura 5. La red semántica

Conviene tener en cuenta a esta altura del desarrollo que se partió de un volumen inicial de 19,532 palabras (extensión del corpus tratado), posteriormente se definió un criterio para determinar las unidades primarias de análisis (frases) y a través del parsing del documento original se obtuvo un universo de 1777 frases. Una forma natural de representación para este corpus fue generar una tabla con cabezal conformado por los N lexicales que aparecen en el corpus y en donde las filas son las frases, en este caso 1777 filas. En cada celda se puede asignar un 1 o 0 ya sea que el lexical este presente o no en la frase. A fines operativos y de optimización se excluyó una parte importante de los lexicales (eliminación de artículos, compactación por lematización, desambiguación, etc.). Esto permitió partir de una tabla inicial de 1777 filas y 160 columnas en donde los lexicales presentes fueron aquellos que mostraban mayor frecuencia de aparición en el documento. A través de la tarea del experto de dominio se llegó finalmente a construir una matriz 1777 filas y 60 columnas. Es de hacer notar que este proceso de reducción está reflejando que si bien el número de frases no se ha modificado, cada columna representa un mayor nivel semántico, lo que indica que estaremos relacionando conceptos cada vez más complejos y enlazando unidades de conocimiento pertenecientes a niveles superiores de abstracción conceptual. Esta será la representación del corpus que servirá como dataset para la clasificación y búsqueda de conocimiento.

#### 4. Clasificación y búsqueda de conocimiento

A principios de los 90 se va incorporando en distintas disciplinas algunos conceptos que han mostrado ser de gran utilidad para el tratamiento de las bases de datos y la extracción de conocimiento. Nos referimos a enfoques tales como Machine Learning, Aprendizaje no Supervisado, Algoritmos, Testeo Redes Neuronales<sup>4</sup>, etc. Estos conceptos y herramientas tienen sus raíces en las Ciencias Estadísticas, en las Ciencias de la Computación y las Ciencias Matemáticas<sup>5</sup>. En sus inicios las aplicaciones fueron principalmente en las áreas de salud, la industria y en menor medida en proyectos comerciales. Con la explosión de Internet, su aplicación se extiende a áreas muy diversas: desde pronósticos comerciales hasta procesos de Marketing Directo, detección de fraudes, asignación de áreas comerciales, optimización de recursos, edición de imágenes, etc.

En nuestra industria existen algunos desarrollos recientes en tratamiento de datos in estructurados<sup>6-7</sup>, las áreas de clasificación de la población por niveles sociales<sup>8</sup>, en los pronósticos de tráfico en sitios<sup>9</sup>, etc.

Para el desarrollo del método que presentamos se utilizaron muchos de los conceptos ya mencionados. En esta etapa se trabajó principalmente sobre el concepto de Clasificación, para ubicar las frases en un mapa que revelara “cercanías” significativas y estableciendo un delicado equilibrio entre la visualización de la información y la pérdida de la misma.

Del universo de posibles clasificadores, optamos por un clasificador de la familia de las redes neuronales, denominado Aprendizaje Competitivo No supervisado, en su versión Self -Organizing Map (SOM). Este tipo de algoritmo permite reducir de una manera drástica la alta dimensionalidad en los datos y a su vez provee cierto grado de abstracción.<sup>10</sup>

Como enfoque alternativo se realizó sobre el mismo dataset un clasificador PCA (Análisis de Componentes Principales), resultando un haz de 60 factores con un poder máximo de explicación de 2.8% en el primer factor. El poder de clasificación del PCA

---

<sup>4</sup> Ver por ejemplo: [http://es.wikipedia.org/wiki/Red\\_neuronal\\_artificial](http://es.wikipedia.org/wiki/Red_neuronal_artificial)

<sup>5</sup> The Elements of Statistical Learning (2003) – T. Hastie; R. Tibshirani y J. Friedman - Stanford University CA

<sup>6</sup> Open Ended Surveys Coding using text Categorization Techniques (2003) – D. Georgetti - ASC Conference 2003

<sup>7</sup> La Semiometrie \_ Essai de Statistique Structurale (2003) – L. Lebart, Jean Francois Steiner - Dunod.

<sup>8</sup> El Big Bang de los Datos (2005) – J.J. Lloret – Primer Congreso de SAIMO

<sup>9</sup> Introducción a las ANN y su Aplicación en la Investigación de Mercados (2000) – Emilio Martínez Ramos – La Investigación en Marketing - AEDEMO

<sup>10</sup> Ver en Apéndice I - The Self-Organizing Map (SOM) by Teuvo Kohonen

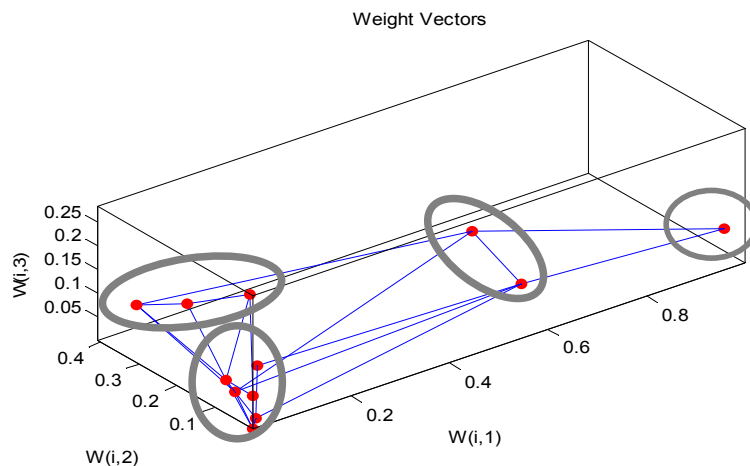
resultó ser apenas mayor al de un proceso totalmente aleatorio, no proveyendo en consecuencia nuevas estructuras relevantes de conocimiento, y mostrando un alto nivel de independencia entre las variables originales.

La entrada al modelo es la matriz  $N \times M$  dicotómica de presencia de la red semántica en el texto, cada celda  $(I,J)$  de la matriz indica si en la frase  $I$  del texto parseado está presente la categoría  $J$  de la red semántica, siendo 1777 la cantidad de frases y 60 las categorías después de las tareas de semantización. El proceso asume entonces las categorías como  $M$  vectores del espacio  $N$  dimensional, analiza su distribución y asigna finalmente a cada frase un nodo de pertenencia.

Gráficamente en el espacio tridimensional, la configuración resultante en una red de 12 nodos con 100 iteraciones por nodo fue la siguiente:

La interpretación semántica del contenido de cada nodo la hará el analista, soportado por las frases  $y$ , por ende, las categorías presentes en dicho nodo y su nivel de pertinencia. En el siguiente cuadro se muestran las características principales de cada nodo obtenido mediante el entrenamiento con la red SOM 3x4\_100, a través de *SOM Toolbox in MATLAB 7.0*.

Asimilando cada nodo como un patrón (*pattern*)<sup>11</sup>



<sup>11</sup> The **pattern** is a [form](#), [template](#), or [model](#) (or, more abstractly, a set of [rules](#)) which can be used to make or to generate things or parts of a thing, especially if the things that are created have enough in common for the underlying pattern to be inferred, in which case the things are said to *exhibit* the pattern (Wikipedia).

Pattern	%	CATEGORÍAS PRESENTES
11	38,0%	AFFECTOS(30%) - COMPROMISO (54%) - DINERO (30%) – EL SUEÑO (40%) - MEJORÍA (35%) - LÍMITES (60%) SOLEDAD (40%) - GRUPO (37%)
1	14,0%	CASA (32%) REF. TEMPORAL (83%) LLORAR SUFRIR (36%)
8	10,1%	INAUGURAL (31%) - SENTIR (64%) - VIDA VIVIR (56%)
3	9,2%	ACONTECIMIENTO (52%) - QUERER (71%)
10	6,4%	LA RACIONALIDAD (47%) LOCURA (38%) BÚSQUEDA DE CURACIÓN (41%) MIEDOS Y TEMORES (47%)
12	6,1%	AUTOPOSICIONAMIENTO (35%) MALESTAR Y ENFERMEDAD (32%)
9	5,1%	PODER (85%)
4	4,1%	ADICCIÓN (17%) DINERO (10%) GENITALIDAD (29%) INAUGURAL (13%) INTENSIDAD (11%) VÍNCULOS CON PARES (5%)
7	3,6%	BUENO MALO (34%) DINERO (10%) INTENSIDAD (30%) SALIR (13%) LLORAR SUFRIR (13%)
2	2,3%	QUERER (22%)
5	0,7%	ADICCIÓN (17%) LÍMITES (10%) PODER (11%)
6	0,5%	LOCURA (13%)

La columna de porcentuales indica el porcentaje de frases que se asignaron a cada patrón, y al lado de cada categoría se indica en forma porcentual, del total de frases en las que está presente la categoría, el porcentual concentrado en el patrón. Por ejemplo, el Patrón 11 acumula el 38% de total de las frases del texto analizado, es decir 1 de cada 3, y además, concentra el 40 % de las frases que hacen referencias a la “soledad” en cualquiera de las formas lexicales en este texto: “la\_soledad”, “estoy\_solo”, me\_siento\_sola”, etc.

Es así que podemos inferir que de un conjunto de palabras originales que formaban discurso en lenguaje natural, ha sido convertido en una estructuración con un mayor grado de abstracción que el inicial permitiendo a) ubicar cada frase afín a un determinado patrón ( es decir identificar a la frase con cualidades de un mayor nivel de abstracción que la frase original) b) establecer relaciones de cercanía y lejanía entre patrones y por ende entre frases y c) poder clasificar con similar estructuración y abstracción una o mas frases nuevas (*newcomers*).

Con fines ilustrativos de la etapa **IV** en el ciclo de la intervención, se incluyen algunas de las conclusiones del experto del dominio (Lic. Martínez, Cap.6, La estructura del Discurso Suicida, México 2007):

“La aplicación de redes neuronales al análisis de contenido del discurso suicida resulta satisfactorio, tanto en su grado de adecuación a la estructura teórica y clínica, como en su nivel de veracidad y de coherencia interna, en esta etapa exploratoria. El algoritmo

está validado en sí mismo como procedimiento matemático libre de errores, pero eso no asegura su utilidad en el campo social o que le transfiera certeza matemática.

El resultado (obtenido) ... ratifica el postulado teórico que sostiene la característica paradójica del discurso suicida, al mostrar en el patrón 11 la coexistencia de las categorías Soledad (40 %) y Grupo (37 %), con una carga similar y en el patrón 10 ... Racionalidad y Locura (47 y 38 % respectivamente).

En el patrón 1 aparecen, también con cargas similares, las categorías Casa (32 %) y Llorar/sufrir (36 %), vinculadas por el origen y la manifestación del malestar instalado en el hablante, ratificado en la Referencia temporal (83 %), categoría tan frecuente a lo largo de todo este discurso. Una interpretación posible de esta vinculación es que, para estos consultantes, en la segunda línea de sus argumentaciones sobre su sufrimiento (14 %), tiene una gran importancia el tiempo vivido y las experiencias acumuladas en los intercambios afectivos con personas significativas y convivientes.

En el patrón 8, con la proximidad de las tres categorías que lo componen, queda expresado el impacto de lo novedoso en el sentir y el vivir de estas personas, dando cuenta del costo de los cambios para la rigidez en la que quedan atrapados frecuentemente.

En el patrón 9 se ubican la mayor cantidad de referencias a la categoría Poder (85 %), que denota capacidad y/o posibilidad, y la categoría vuelve a aparecer en un patrón residual, el 5, junto con Adicción –que no resulta significativa en este corpus discursivo- y Límites, alusivo a limitaciones e incapacidades. Esta correspondencia vuelve a mostrar, por un lado el aspecto paradójico de las enunciaciones (Poder / Límites), y por otro que las percepciones de capacidad de estos consultantes están asociadas en una misma dirección.

...

Este tipo de intervención se torna más relevante si se tiene en cuenta que en el patrón 12 se vinculan sólo dos categorías: Autoposicionamiento y Malestar y Enfermedad, lo que denota la percepción predominante de estos integrantes del grupo con respecto al momento presente de sus vidas.

Un ejemplo claro de coherencia interna ... es la presencia en el patrón 10, de las categorías Búsqueda de curación y Hablar. Todas las vías de recuperación ensayadas por estos consultantes, insertos en una cultura urbana, están vinculadas a la acción de hablar (tratamiento, terapia, terapeutas, psicólogo/a, psiquiatra).

... (el método permitió) una medición de la ubicación contextual de la palabra “locura”.

... la categoría Locura aparece con la segunda menor carga (38 %) en el patrón 10 (6,4 %), y en un patrón residual, con una carga poco significativa (13 %). Se puede

inferir que no siempre las prioridades del marco teórico del investigador, coinciden con las prioridades de las emergencias discursivas que debe analizar.

En esta misma línea se puede observar que, en un cuadro que resulta develador de la estructura del discurso suicida, las categorías Morir y Suicidio y parasuicidios ... no resultan significativas para la estructura discursiva en estudio, aunque sí lo sean para el marco teórico de quien investiga.

... queda por evaluar la precisión diferencial de los resultados de la aplicación de la red neuronal, si el discurso estuviera clasificado por nivel de riesgo, alto, moderado o bajo, medido por ISO 30.

...

Por último esta experiencia piloto parece demostrar la utilidad del análisis discursivo por este método, en las manifestaciones tempranas o fallidas de la secuencia suicida, antes que hacerlo sobre un corpus de notas suicidas.”

## 5. Conclusiones

El proceso de incorporar valor agregado a los datos, de forma tal de convertir una informe masa inicial en una estructura de conocimiento útil, es sin duda un reto formidable para nuestra Industria.

Este desafío requiere movilizar no sólo recursos económicos sino también capacidades intelectuales multidisciplinarias, y así evitar el serio riesgo de desaparecer como actividad económica al ser depredada por otros actores y disciplinas que desde la óptica de nuestros usuarios tienen una valorización más positiva en términos de actualización tecnológica, eficacia y modernidad.

El presente ensayo muestra las posibilidades de una ingeniería interdisciplinaria, donde se trata de organizar los grandes volúmenes de datos “multifuentes”, de manera de extraer un mayor valor semántico que el que se obtiene de un barrido bidimensional cartesiano. Este valor agregado permite ver los datos en una forma de imagen o de mapa, que entendemos da una descripción del fenómeno, cualitativamente mas rica y con mayor poder de síntesis, que el utilizado actualmente.

Este desarrollo esta orientado – entre otras - a las siguientes aplicaciones dentro de nuestra Industria:

- ♦ Organización de grandes volúmenes *multifuentes*
- ♦ Procedimiento en datos observacionales cualitativos
- ♦ Clasificación y síntesis de bibliotecas temáticas, reports, papers, diarios autoadministrados, blogs, sitios web, etc.
- ♦ Tratamiento de fenómenos temporales y visualización sintética de los eventos
- ♦ Tratamiento de datos provenientes de aplicaciones en Web 2.0
- ♦ Desarrollo de redes de Open Source Thinking
- ♦ Organización y mapeo de monitores de noticias sobre un tema o concepto específico. Detección de tendencias y hechos singulares
- ♦ Tratamiento de entrevistas en profundidad
- ♦ Codificación de preguntas abiertas en forma quasi-automática
- ♦ Estructuración de conjuntos de grupos de discusión

## APÉNDICE

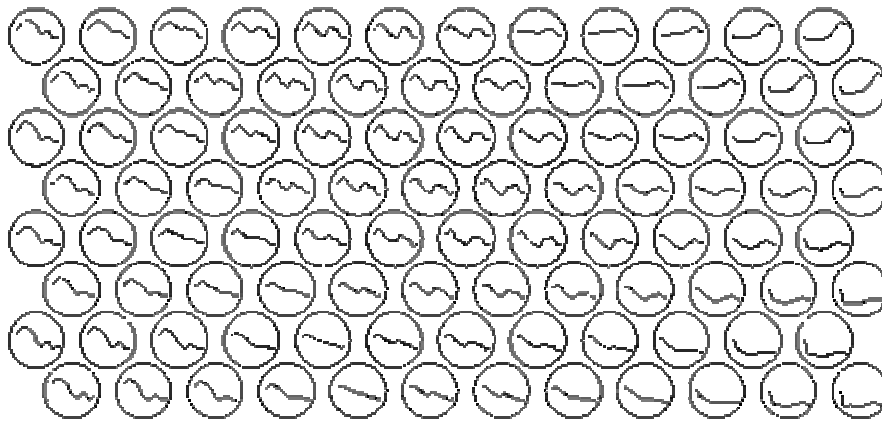
### THE SELF-ORGANIZING MAP (SOM) *por* TEUVO KOHONEN

*extraído de <http://www.cis.hut.fi/projects/somtoolbox/>*

#### Introduction

The SOM is a new, effective software tool for the visualization of high-dimensional data. It converts complex, nonlinear statistical relationships between high-dimensional data items into simple geometric relationships on a low-dimensional display. As it thereby compresses information while preserving the most important topological and metric relationships of the primary data items on the display, it may also be thought to produce some kind of abstractions. These two aspects, visualization and abstraction, can be utilized in a number of ways in complex tasks such as process analysis, machine perception, control, and communication.

The SOM usually consists of a two-dimensional regular grid of nodes. A model of some observation is associated with each node (cf. Fig. 1).



*Figure 1:* In this exemplary application, each processing element in the hexagonal grid holds a model of a short-time spectrum of natural speech (Finnish). Notice that neighboring models are mutually similar.

The SOM algorithm computes the models so that they optimally describe the domain of (discrete or continuously distributed) observations.

The models are organized into a meaningful two-dimensional order in which similar models are closer to each other in the grid than the more dissimilar ones. In this sense the SOM is a similarity graph, and a clustering diagram, too. Its computation is a nonparametric, recursive regression process.

### The incremental-learning SOM algorithm

Regression of an ordered set of model vectors  $\mathbf{m}_i \in \mathbb{R}^n$  into the space of observation vectors  $\mathbf{x} \in \mathbb{R}^n$  is often made by the following process:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{c(x),i}(\mathbf{x}(t) - \mathbf{m}_i(t)), \quad (1)$$

where  $t$  is the sample index of the regression step, whereby the regression is performed recursively for each presentation of a sample of  $\mathbf{x}$ . Index  $c$  ("winner") is defined by the condition

$$\|\mathbf{x}(t) - \mathbf{m}_c(t)\| \leq \|\mathbf{x}(t) - \mathbf{m}_i(t)\| \quad \forall i. \quad (2)$$

Here  $h_{c(x),i}$  is called the *neighborhood function*, and it is like a smoothing kernel that is time-variable and its location depends on condition in equation (2). It is a decreasing function of the distance between the the  $i$ th and  $c$ th models on the map grid.

The neighborhood function is often taken to be the Gaussian

$$h_{c(x),i} = \alpha(t) \exp\left(-\frac{\|\mathbf{r}_i - \mathbf{r}_c\|^2}{2\sigma^2(t)}\right), \quad (3)$$

where  $0 < \alpha(t) < 1$  is the learning-rate factor, which decreases monotonically with the regression steps,  $\mathbf{r}_i \in \mathbb{R}^2$  and  $\mathbf{r}_c \in \mathbb{R}^2$  are the vectorial locations in the display grid, and  $\sigma(t)$  corresponds to the width of the neighborhood function, which is also decreasing monotonically with the regression steps.

A simpler definition of  $h_{c(x),i}$  is the following:  $h_{c(x),i} = \alpha(t)$  if  $\|\mathbf{r}_i - \mathbf{r}_c\|$  is smaller than a given radius around node  $c$  (whereby this radius is a monotonically decreasing function of the regression steps, too), but otherwise  $h_{c(x),i} = 0$ . In this case we shall call the set of nodes that lie within the given radius the *neighborhood set*  $N_c$ .

Due to the many stages in the development of the SOM method and its variations, there is often useless historical ballast in the computations.

For instance, one old ineffective principle is random initialization of the model vectors  $\mathbf{m}_i$ . Random initialization was originally used to show that there exists a strong self-organizing tendency in the SOM, so that the order can even emerge when starting from a completely unordered state, but this need not be demonstrated every time. On the contrary, if the initial values for the model vectors are selected as a regular array of vectorial values that lie on the subspace spanned by the eigenvectors corresponding to the two largest principal components of input data, computation of the SOM can be made orders of magnitude faster, since (i) the SOM is then already approximately organized in the beginning, (ii) one can start with a narrower neighborhood function and smaller learning-rate factor.

Many computational aspects like this have been discussed in the software package SOM\_PAK [1], as well as the book [2].

### The batch version of the SOM

Another remark concerns faster algorithms. The incremental regression process defined by equations (1) and (2) can often be replaced by the following batch computation version which, especially with Matlab functions, is significantly faster.

If all observation samples  $\mathbf{x}(t), t = 1 \dots N$  are available prior to computations, they can be applied as a batch in the regression, whereby the following computational scheme can be used [2]:

- Initialize the model vectors  $\mathbf{m}_i$ .
- For each map unit  $i$ , collect a list of all those observation samples  $\mathbf{x}(t)$ , whose most similar model vector belongs to the neighborhood set  $N_i$  of node  $i$ .
- Take for each new model vector the mean over the respective list.
- Repeat from step 2 a few times.

Notice that steps 2 and 3 need less memory if at step 2 we only make lists of the observation samples  $\mathbf{x}(t)$  at those units that have been selected for winner, and at step 3 we form the mean *over the union of the lists* that belong to the neighborhood set  $N_i$  of unit  $i$ .

### Further remarks

Finally it should be taken into account that the purpose of the SOM is usually visualization of data spaces. For an improved quality (isotropy) of the display it is advisable to select the grid of the SOM units as hexagonal; the reason is similar as when using a hexagonal halftone raster for images.

The above algorithms can often be generalized by defining various generalized matching criteria.

The following categories of similarity graphs, computed by the SOM, have already been used in many practical applications:

- State diagrams for processes and machines
- Data mining applications: similarity graphs for statistical tables and full-text document collections

A list of research papers from very different application areas of the SOM and its variations is available in the Internet at the WWW address <http://www.cis.hut.fi/research/som-bibl/>.

## References

[1] Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J. (1996). SOM\_PAK: The self-organizing map program package. Report A31. Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland. Also available in the Internet at the address [http://www.cis.hut.fi/research/som\\_lvq\\_pak.shtml](http://www.cis.hut.fi/research/som_lvq_pak.shtml).

[2] Kohonen, T. (1995). *Self-Organizing Maps*. Series in Information Sciences, Vol. 30. Springer, Heidelberg. Second ed. 1997.

## BIBLIOGRAFIA

- Anand Reshma Mining blogs: pertinent Chat 2007 In breief January/ February AQR Publication
- Balbi Simona, Di Meglio Emilio A Text Mining Strategy. based on Local Contexts of Words 2004 Dip. di Matematica e Statistica – Università “Federico II” di Napoli
- De Beaugrande, Robert y Wolfgang Dressler. Introducción a la lingüística del texto. 1997 Editorial Ariel. Barcelona.
- Del Brío, Bonifacio M. y Alfredo Sanz Molina.. Redes Neuronales y Sistemas Difusos. 2001 Alfaomega Grupo Editor. México.
- Georgetti D. - Open Ended Surveys Coding using text Categorization Techniques 2003 – ASC Conference 2003
- Hastie T.; Tibshirani R. y Friedman J. The Elements of Statistical Learning 2003 — Stanford University CA
- Hertz J., Krogh A. Palmer R Introduction to Theory of Neural Computation 1991 Santa Fe Institute Studies in the Science of Complexity – Addison Wesley Publishing Company
- Kohonen Teuvo - The Self-Organization and Associative Memory 1989 Berlin Springer - Verlag
- Lebart L. ,. Steiner J.F - La Semiometrie \_ Essai de Statistique Structurale 2003 — Dunod.
- Lloret J.J. El Big Bang de los Datos 2005 (Parte I) — Primer Congreso de SAIMO
- Manning, Cristopher y Hinrich Schutze. Foundations of Statistical Natural Language Processing. 2002 The MIT Press. Massachusetts.
- Martinez Ramos Emilio Introducción a las ANN y su Aplicación en la Investigación de Mercados 2000 — La Investigación en Marketing - AEDEMO
- Sebastiani Fabrizio Text Mining and its Applications WIT Press, Southampton, UK, 2005,
- Sebastiani, Fabrizio. Machine Learning in Automated Text Categorization. 2002 ACM Computing Surveys, Vol 34, No.1,pp 1-47.

- Schmidt Aplicaciones de redes neuronales para analizar Focus Groups (2001) Marcus, Qualitative Market Research. Journal
- Sperber, Dan y Deirdre Wilson. La Relevancia: comunicación y procesos cognitivos. Editorial 1994 Visor. Madrid.
- Vilarroya Oscar – La disolución de la mente – 2002 Metatemas TUSQUETS editores