UNA METRICA PARA LA ASOCIACION DE PALABRAS: JUNG NOT DEAD

Martín Elías Costa

Tesis de Licenciatura en Ciencias Físicas

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Marzo 2008

TEMA: Neurociencia Cognitiva

ALUMNO: LU N° : 29/03

LUGAR DE TRABAJO: Laboratorio de Neurociencia Integrativa

DIRECTOR DEL TRABAJO: Mariano Sigman

CODIRECTOR o COLABORADOR: Flavia Bonomo

FECHA DE INICIACION: Abril 2007

FECHA DE FINALIZACION: Marzo 2008

FECHA DE EXAMEN: 31 marzo 2008

INFORME FINAL APROBADO POR:

Autor	Jurado
Director	Jurado
Profesor de Tesis de Licenciatura	Jurado

Capítulo 1 Introducción

El estudio del lenguaje provee una ventana de acceso a la forma en que nuestros cerebros almacenan, organizan, recuperan y manipulan información. Durante el discurso fluido – el habla o la escritura - estas tareas son coordinadas para concatenar y combinar ideas con flexibilidad, lo cual permite generar un repertorio infinito de conceptos coordinadamente y entendernos. Para que haya comunicación, el mensaje debe, luego, ser decodificado por otro cerebro. En este proceso el receptor acompaña el recorrido de conceptos y analiza las relaciones entre los mismos (de ahí la expresión: *¿me seguís?*). Del hecho de que la comunicación se

produzca (de que a geinchi se entienda) se infiere que la forma en que están asociados los significados en distintos sujetos es similar.

Otro fenómeno de la vida cotidiana que sugiere que existe un orden en el espacio de las palabras es la evocación espontánea de una palabra por otra. Cuando escuchamos o leemos una palabra se nos ocurren "naturalmente" otras. Por ejemplo en este momento escribo la palabra VAPOR y se me ocurren: BARCO, CHIMENEA, MOTOR, MAQUINA, TREN, etc. Algunas asociaciones, de hecho, son casi inevitables. Frente a la palabra BLANCO, casi todo el mundo, casi todas las veces piensa en la palabra NEGRO. Presentado con la palabra PLATO, el número de palabras asociadas con alta frecuencia es más grande: SOPA, CUCHILLO, TENEDOR… El primero en aprovechar esto para intentar sondear los recovecos de la mente fue el psicólogo austriaco Carl Jung.

Jung tenía la idea de que los sentimientos, ideas, experiencias, etc. estaban agrupados en el inconsciente por medio de asociaciones. Llamó a estos grupos complejos e ideó un método para diagnosticar patologías relacionadas con cada complejo utilizando lo que ahora se conoce como el método de asociaciones de palabras. Presentaba a sus pacientes una serie de palabras estímulo y les pedía que contestaran lo más rápido posible la primera palabra que se les viniera a la cabeza. Analizaba los

tiempos de respuesta y las palabras asociadas para hacer su diagnóstico. Jung, además, asumía que este método establecía una ventana al diagnostico y por ende a la terapia. Nosotros usaremos el mismo método pero con miras a tratar de entender como es la organización del espacio de significados (léxico).

Este tipo de asociación entre memorias es ubicuo a través de las distintas representaciones sensoriales. Una imagen, despierta otra imagen, un sonido, otro; existiendo también evocaciones cruzadas (un olor dispara una imagen) sugiriendo que la matriz de asociaciones y la categorización se da en distintos terrenos. Uno asocia una manzana con una pera porque ambas son frutas, así como la nieve y la leche porque ambas son del mismo color, o porque existe alguna otra etiqueta morfológica o emocional. Un posible modelo para las bases neurales de la memoria asociativa es una versión modificada del modelo de Ising que se conoce con el nombre de redes de Hopfield (para una aplicación de estas redes a la resolución de problemas de Sudoku ver el trabajo de Hopfield, 2006). En él, cada neurona es una unidad que puede estar en dos estados (espín up y espín down) y que se conecta con otras mediante una matriz de interacción. Las memorias están representadas por aquellos patrones de activación de las unidades que sean mínimos locales de energía. Si se deja evolucionar al sistema desde un estado arbitrario éste converge al mínimo de energía más cercano (en el espacio de activaciones). La

interpretación de esto, en el contexto que nos ocupa es la siguiente: la palabra que se presenta es el estado de activación inicial (input) y el sistema evoluciona hacia la palabra más "cercana" en algún sentido que resulta difícil formalizar.

Intuitivamente, palabras que se encuentran cerca unas de otras son palabras que pertenecen al mismo contexto. Decidir si dos palabras pertenecen o no al mismo contexto ha sido tradicionalmente tarea de los lingüistas. Algunos de ellos se han tomado el trabajo de armar grandes bases de datos con los distintos tipos de relaciones entre palabras. Por ejemplo, Miller, Fellbaum y colegas, generaron el sitio Wordnet, una red semántica en ingles donde se registran y clasifican distintos tipos de relaciones semánticas. La posibilidad de tener acceso a este tipo de datos en forma automatizada motivo una gran cantidad de estudios cuantitativos sobre la organización del léxico. Por ejemplo, Sigman y Cecchi (2003) estudiaron la estructura de la red Wordnet y mostraron, entre otras cosas, que la existencia de palabras con más de un significado le confiere al léxico características de red small world.

Si bien este tipo de redes semánticas no nos permite evaluar si una palabra esta más asociada con una que otra, sí nos puede servir para ver si lo que nosotros definimos como contexto tiene un correlato en nuestros cerebros. Hay numerosos experimentos que

muestran que sí. Las variantes son muchas pero la idea esencial siempre es la siguiente: se le da al sujeto una tarea que involucre dos o más palabras y se mide el tiempo que tarda en realizarla. Se ve que ese tiempo es más corto si las palabras pertenecen al mismo contexto. Para poner un ejemplo concreto tomemos los experimentos de Meyer y Schvaneveldt (1971). Ellos les pedían a los sujetos que identificaran si las palabras que aparecían en una pantalla eran efectivamente palabras (en inglés) o tiras de caracteres. Lo que observaron fue que los tiempos de respuesta al identificar una palabra (p Ej. DOCTOR) eran menores si la palabra había sido precedida por una relacionada semánticamente. Es decir, la secuencia NURSE→DOCTOR era reconocida más fácilmente que BUTTER→DOCTOR.

Todas estas cosas, desde los experimentos de priming hasta la experiencia cotidiana, pasando por la idea de los complejos de Jung y las memorias como mínimos de energía en una red de Hopfield, parecen sugerir que es posible asociar a cada par de palabras una cantidad que indique en que grado pertenecen al mismo contexto. Podemos, como ejercicio mental, interpretar esa cantidad como una distancia entre palabras. Esto abre las puertas a un nuevo marco en el cual podemos pensar al léxico como un paisaje. No es difícil imaginar un mundo de palabras en el que si caminando vemos la palabra MANZANA esperamos encontrar cerca PERA, DURAZNO

y otras frutas. Siguiendo estará BANANA que forma una esquina, a la vuelta de la cual está MONO.

A pesar de que la existencia de dicho mundo resulta intuitiva y está respaldada tanto por experimentos como por teoría, definir una distancia entre conceptos resulta complicado. En este trabajo, proponemos una manera posible. La primera pregunta que surge, lógicamente es ¿Cómo representar el espacio de significados? ¿Qué herramienta matemática nos permite capturar mejor sus características?

Estas preguntas nos empujan a mirar nuevamente en los dominios de filósofos y lingüistas. Las escuelas holísticas (Quine 1969), proponen una visión en la cual el significado no es algo inmutable que relaciona un concepto con su contraparte en el mundo real (objeto) sino que depende del contexto. Es decir que un concepto es, también, todas sus relaciones con otros conceptos. Ferdinand de Saussure, cuando define sus relaciones asociativas, escribe: "Un término dado es como el centro de una constelación, el punto donde convergen otros términos coordinados cuya suma es indefinida." Inclusive, la forma más tradicional de organizar los conceptos, el diccionario, sugiere la misma idea. Allí, las palabras aparecen definidas en términos de otras con las que están, por construcción, asociadas. Cada palabra tiene, alrededor, un núcleo de otras palabras que forma su entorno de asociación. Changizi (2007),

por ejemplo, utiliza una red generada a partir de cuales son las palabras que aparecen en una dada definición del diccionario para estudiar la estructura jerárquica del lenguaje.

Todo esto sugiere que las palabras integran un espacio donde algunos de sus elementos están relacionados. Una buena forma de representar esto es usar un grafo, donde cada palabra sea un nodo y las relaciones entre ellas sean las aristas. Esta representación es útil a la hora de analizar cuantitativamente la organización del léxico pues existen numerosas herramientas para caracterizar la estructura de un grafo así como también la dinámica en el espacio que define.

Una parte importante de este trabajo se centra en definir, de manera operativa, cuales son las palabras que están conectadas y cuan fuerte es su relación; es decir, cual es la conformación del grafo. Luego, nos focalizaremos en algunos aspectos de la dinámica con el fin de entender los mecanismos generativos de las secuencias de asociaciones libres.

Capítulo 2 Definición de distancia

Siguiendo la idea de que lo que intentamos capturar es el grado de pertenencia de dos pares de palabras a un mismo contexto, intentaremos definir una métrica a partir de la estructura estadística de las palabras en textos. El concepto básico detrás de esta métrica es que dos palabras están próximas entre si, si aparecen juntas (en algún sentido que resta definir) muy seguido. Esto sugiere una relación directa entre métrica y probabilidad: palabras con altas probabilidad de ocurrir en el mismo contexto deberían estar a una distancia cercana, es decir, grosso modo: $d(x,y) \approx \frac{1}{p(x,y)}$. En las siguientes secciones nos avocaremos a formalizar esta idea utilizando nociones de probabilidades y a estudiar como estimar dichas probabilidades a partir de grandes corpus de texto.

2.1- Probabilidades condicionales y frecuencias relativas.

La primera noción importante, a fin de establecer una métrica entre pares de palabras es la de probabilidades condicionales y conjuntas. Si las probabilidades de que dos eventos A y B ocurran están dadas por P(A) y P(B) respectivamente; y la probabilidad de que ambos ocurran es $P(A \cap B)$, se define la probabilidad condicional de que ocurra A dado que B ha ocurrido como:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \textbf{(2.1)}$$

Vale la pena hacer un pequeño ejemplo para comprender el significado de probabilidad condicional. Supongamos que tenemos un mazo de truco de cuarenta cartas y que el evento A es sacar un ancho de espadas y el B es simplemente sacar una espada. La probabilidad de sacar el ancho dado que sabemos que sacamos una espada es, intuitivamente, un décimo. Para hacer la cuenta, usamos la **(2.1)**: la probabilidad de sacar un ancho de espadas y que sea de espadas es simplemente $\frac{1}{40}$ mientras que la probabilidad de sacar

una espada es $\frac{1}{4}$. Dividiendo obtenemos el mismo resultado que nos parecía intuitivo. Otra cosa que este ejemplo nos permite observar inmediatamente es que en general: $P(A \mid B) \neq P(B \mid A)$. Específicamente, la probabilidad de sacar el ancho dado que sacamos una espada es un décimo pero la probabilidad sacar una espada dado que sacamos el ancho es, evidentemente uno.

En ciertas circunstancias (como será de hecho nuestro caso) uno no tiene acceso a las probabilidades sino al número de eventos. Esto, sin embargo, no resulta un inconveniente ya que la probabilidad condicional puede estimarse a partir del conteo de eventos. Es posible reescribir la **(2.1)** a partir del número de eventos compatibles con A (o B) $N(A)$ ($N(B)$) y el número total de eventos Ω:

$$P(A \mid B) = \frac{\frac{N(A \cap B)}{\Omega}}{\frac{N(B)}{\Omega}} \qquad \textbf{(2.2)}$$

que no depende, verdaderamente, del factor de normalización Ω. La ecuación **(2.2)** nos permite estimar la probabilidad condicional contando la fracción de eventos en los que ocurren tanto A como B respecto de en los que ocurre solo B.

A modo de ejemplo, supongamos que contamos con un corpus de texto, a partir del cual podemos estimar las probabilidades de ocurrencia de distintas palabras (algunas como blanco son mas

frecuentes que otras, como xilofón) y la probabilidad de co-ocurrencia de dos palabras. Esta co-ocurrencia puede determinarse, según la norma, en la misma frase, en el mismo párrafo, en el mismo texto, en la misma pagina Web etc… y mas abajo discutiremos distintas alternativas. En tal caso, la probabilidad condicional $P(PERRO \mid GATO)$ establece una medida de correlación (o, simplemente de relación entre dos palabras). La inversa de esta medida es un estimativo de la distancia y, por lo tanto, podemos definir la distancia entre dos palabras PERRO y GATO, como:

$$D(PERRO, GATO) = \frac{1}{P(PERRO \mid GATO)} \qquad \textbf{(2.3)}$$

Dado que en ciertas circunstancias (como en las paginas Web) uno no tiene acceso a la integridad del corpus y por lo tanto puede contar eventos pero no probabilidades, esta distancia puede reescribirse de manera equivalente, -según la deducción simple mostrada anteriormente - reescribiendo en término del número de veces que aparece una dada palabra en una unidad de texto (p ej. una oración, párrafo, capítulo, página Web, etc.):

$$D(PERRO, GATO) = \frac{N(GATO)}{N(PERRO \cap GATO)} \qquad \textbf{(2.4)}$$

Una lista de palabras y una distancia entre ellas establecen precisamente un grafo donde los nodos son las palabras y las aristas

tienen distintos pesos según la distancia entre cada par de palabras. A estos grafos (en los cuales las aristas tienen peso) se los llama, grafos pesados. Una buena forma de visualizar esta estructura es a través de una matriz, donde el elemente D_{ij} corresponde a la distancia entre la palabra *i* y la *j*.

En resumen: a partir de un corpus uno puede contar cuantas veces dos palabras aparecen de manera conjunta (debidamente normalizado por la probabilidad de ocurrencia de cada una), a partir de eso definir una métrica. Esto convierte al conjunto de todas las palabras en un grafo pesado.

2.2- Artefactos en la norma Google®

Implementada la estrategia, el siguiente paso era determinar el corpus de texto a partir del cual generar las probabilidades conjuntas. Nuestro primer intento fue utilizar la cantidad de resultados que devuelve un buscador Web (p. ej. Google®), en tal circunstancia, bastaba con contar el numero de paginas donde aparece la palabra *i*, la palabra *j*, y, de manera conjunta, las palabras *i* y *j*. Elegimos inicialmente google dado que indexa un número inmenso de páginas lo cual debía garantizar una buena convergencia de estos conteos. Escribimos un programa en Perl que automatiza las consultas al buscador y recupera el número de

resultados obtenidos a partir del código HTML. Algunas pruebas rápidas nos permiten poner en evidencia que el número que brindan estos buscadores no puede ser simplemente la cantidad de páginas Web en las que aparece una dada palabra. Por ejemplo, si buscamos las frecuencias de aparición de todas las palabras de un diccionario y hacemos un histograma de las mismas vemos que hay picos (ver Figura 2.1) alrededor de 1.800.000 y 2.700.000. Esto es completamente inesperado y contradice resultados bien establecidos sobre la frecuencia de aparición de palabras en textos como la ley de Zipf que establece que la distribución de frecuencias sigue una ley de potencias. Por otro lado, cuando hacemos una búsqueda de dos palabras la cantidad de resultados que obtengamos dependerá del orden en el que las hayamos escrito (p. ej. La búsqueda GATO+PERRO devuelve 562.000 resultados mientras que PERRO+GATO devuelve 1.330.000) Si bien estas asimetrías pueden darse en ciertos algoritmos de conteo en texto (por ejemplo, la probabilidad de que ocurran en la misma frase gato seguido de blanco es mayor que blanco seguido de gato), en el caso de google uno no tiene acceso a cuales son los algoritmos que generan estas asimetrías. Por estas dificultades, decidimos abandonar esta forma sencilla de búsqueda de probabilidades que utiliza algoritmos ya establecidos y que aprovecha una base de datos inmensa. Nos abordamos entonces a generar nuestros propios algoritmos para

contar eficientemente la ocurrencia y co-ocurrencia de palabras en grandes corpus de texto.

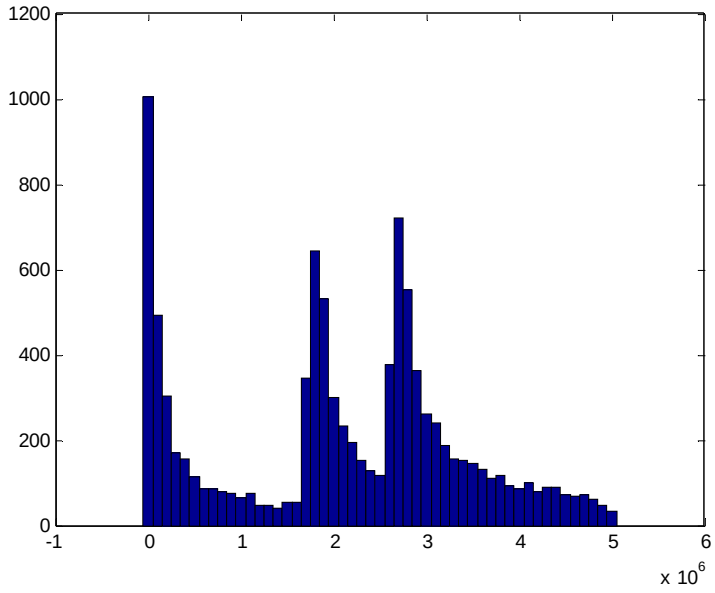

Figura 2.1: Histograma de las frecuencias de aparición en Google de todas las palabras del diccionario.

2.3- Norma de estadística en textos.

Antes de poder contar las ocurrencias y co-ocurrencias necesitamos tener una gran base de textos en el cual hacerlo. En ésta sección describimos el armado y organización de dicha base así como también los algoritmos de conteo que nos permiten pasar de texto a la matriz de distancia.

La primera cuestión que surge es de donde sacar los textos. Existen varios sitios en Internet que almacenan libros sin copyright y pueden ser usados libremente. Con otro programa sencillo, que también escribimos en Perl, bajamos textos en español del sitio Web del proyecto *Gutemberg* así como también notas del diario *La Nacion*, con permiso de sus autores. La tarea de homogenizar los textos provenientes de fuentes distintas no es fácil. Cada sitio tiene su propia forma de catalogar los textos, darles formato, algunos vienen con encabezados en otros idiomas, etc. Muchas veces, están también embebidos en código fuente (HTML, Java Script, PHP, etc.). Para poder seguir adelante y efectuar un conteo correcto

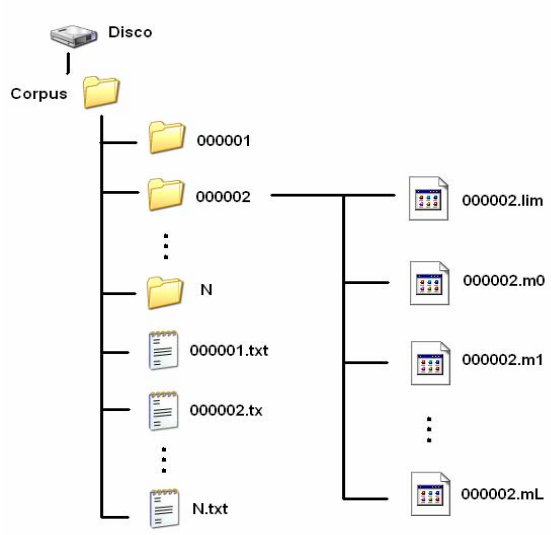

Figura 2.2: Esquema de almacenamiento de los textos. El archivo *.lim contiene el texto sin caracteres extraños y los *.m*l* tienen los conteos parciales de palabras.

tenemos que limpiar los textos. Es para estas tareas que Perl resulta particularmente útil. Haciendo uso de expresiones regulares (ver porque perl) podemos eliminar, en forma bastante general, todos los elementos que dificulten el posterior conteo de palabras (incluyendo signos de puntuación). Cada texto es indexado y almacenado siguiendo el esquema que puede verse en la Figura 2.2. Los archivos .txt contienen los textos originales y los .lim las versiones filtradas sin caracteres extraños. Luego efectuamos un conteo parcial que se guarda en los archivos *.m*l*. Cada uno de esos contiene todos los pares distintos de palabras que aparecen en el texto a distancia *l*; *l* es la cantidad de palabras que las separan. En la Figura 2.3 se puede ver como una frase corta se transformaría en los archivos frase.m0, frase.m1 y frase.m2. El archivo *.m0 contiene la frecuencia de aparición de cada palabra para todas las palabras en ese texto. Si lo que queremos es contar la cantidad de veces que dos palabras aparecen juntas a menos de M palabras de separación debemos sumar sobre todos los archivos *.ml con l **menor o igual que** M. Esto es porque lo que contamos antes es la cantidad de veces que aparecen un par de palabras separadas por una distancia de **exactamente** l. Si queremos computar estas co-ocurrencias para una lista de palabras dada, podemos asignarle a cada una un índice y calcular los elementos de una matriz de co-ocurrencias definida por:

$$C_{ij} = \sum_{l=1}^{M} N_l(i,j) \qquad \textbf{(2.5)}$$

Donde los $N_l(i,j)$ se obtienen directamente de los archivos *.m*l*

Para completar la definición de distancia falta elegir esta

Figura 2.3: Conversión de los textos en archivos *.m*i*

especie de longitud de correlación *M*. Tomaremos como *M* el largo típico de una frase (~ 9 palabras). Resulta interesante, pero queda por fuera de este trabajo, estudiar como cambian las propiedades de la distancia así definida con el parámetro *M*. Para M = 1, estaremos tomando solamente palabras contiguas. Un análisis sobre la estructura de un grafo generado de esta manera puede encontrarse en un trabajo de Ferrer i Cancho y Solé (2001) y más en general sobre la estadística de textos para relaciones contiguas en el trabajo de Shannon (1963) en el que sienta las bases de la teoría de la información.

Elegido el M, podemos (a partir de la matriz de co-ocurrencias y la frecuencia total de cada palabra) construir la matriz de distancias usando la ecuación **(2.4)**.

Acá urge una pequeña aclaración. En el ejemplo sencillo del mazo de truco, vimos que en general $P(A \mid B) \neq P(B \mid A)$ y por lo tanto nuestra matriz de distancias heredará esta propiedad (será asimétrica). Es decir que no estamos ante la definición de una función que cumpla con todos los axiomas que definen una distancia. Esto no debe inquietarnos ya que ninguno de los análisis que hagamos dependerá de que se cumpla este axioma. Más aún, esta asimetría parece ser una característica intrínseca del proceso de transición en un experimento de asociación libre. Esto puede entenderse fácilmente con un ejemplo: las palabras NEGRO y GATO co-ocurren seguido, sin embargo, NEGRO es mucho más frecuente que GATO. Es decir que la transición GATO → NEGRO será más probable que NEGRO → GATO simplemente por el hecho de que NEGRO es una palabra mucho mas conectada. Con esto damos fin al breve paréntesis sobre la asimetría de la norma.

Hemos definido, entonces nuestra matriz de distancias y estamos listos para empezar a medir, ¿o no? Un problema que surge inmediatamente, debido al tamaño finito del corpus, es que varios pares de palabras no aparecen nunca juntas; dando como resultado distancias infinitas. La forma de suavizar dichos infinitos es a lo que nos dedicaremos en la próxima sección.

2.4- El problema de los infinitos: Algoritmo de Dijkstra

El algoritmo de Dijkstra fue inventado en 1959 por el computador científico holandés Edsger Dijkstra y resuelve el problema de encontrar el camino mínimo entre vértices de un grafo pesado, con pesos no negativos. Antes de entrar en una descripción detallada del algoritmo, veamos como es que esta idea nos permite renormalizar los infinitos que aparecen en la matriz de distancias.

Supongamos que después de computar las frecuencias relativas entre palabras y obtener la matriz de distancias, D(PERRO, ANIMAL) y D(GATO, ANIMAL) están definidas pero D(PERRO, GATO) no. Estaríamos tentados, por consiguiente, de reemplazar la distancia infinita entre PERRO y GATO por la suma de las distancias de ir de PERRO a ANIMAL y luego de ANIMAL a GATO. Siguiendo esta idea, construimos un nuevo grafo en el que la distancia entre dos palabras es la que se obtiene de recorrer el camino más corto entre ellas dos en el grafo original (el definido en la sección anterior). Esto no solo resuelve el problema de los infinitos (suponiendo que el grafo original es conexo) sino que tiene sentido dentro del contexto de la teoría asociativa de la memoria. En un experimento de asociación libre uno espera observar la transición espontánea entre un concepto y otro formando una

cadena y es posible que no todos los eslabones de esa cadena tengan acceso a conciencia. Considerar caminos en el grafo original de orden mayor a uno es también una forma de capturar ese comportamiento.

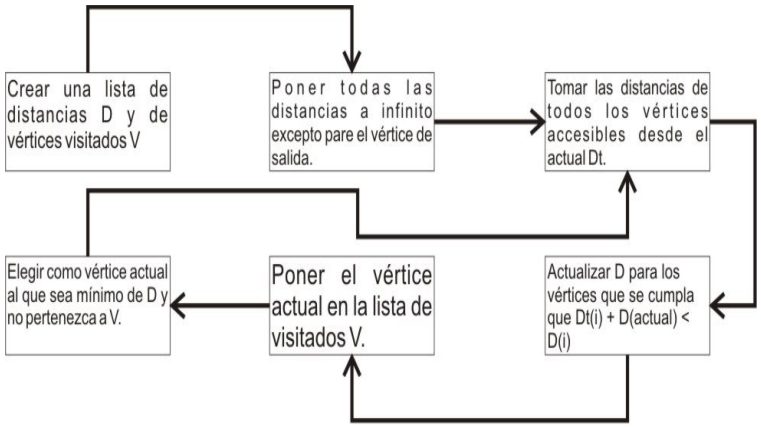

Figura 2.4: Esquema de funcionamiento del algoritmo de Dijkstra.

Ahora sí, dediquemos algunas líneas a estudiar el funcionamiento del algoritmo y como implementarlo en este problema en particular. El mismo pertenece a una clase de algoritmos que se conocen como *golosos (greedy)* porque en cada paso eligen la opción mínima local con la esperanza de obtener finalmente un mínimo global. En general los algoritmos golosos no son capaces de encontrar óptimos globales (especialmente en problemas NP-completos) pero existen algunos problemas en los que es posible demostrar que encuentran siempre la solución global

y en esos casos son algoritmos muy eficientes. Éste es uno de esos casos. En la Figura 2.4 se puede ver la secuencia de pasos que permite resolver el problema. Los mismos pasos se repiten partiendo de cada uno de los vértices del grafo para obtener una nueva matriz *dijkterizada*.

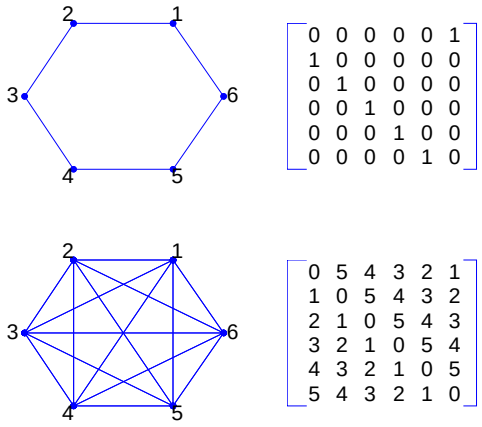

Figura 2.5: Grafo circular dirigido y su representación matricial antes y después de aplicar el algoritmo de Dijkstra.

En síntesis, el algoritmo de Dijkstra puede utilizarse como un operador de grafos. Si el grafo original es conexo (lo cual quiere decir que existe un camino que une dos nodos cualesquiera del grafo) entonces el grafo resultante luego de aplicar el "operador" de Dijkstra tiene las siguientes características: 1) Todos los elementos están conectados entre si (por un peso proporcional al camino mínimo del grafo inicial) y 2) si existía una arista entre dos nodos

i,j, entonces la arista del grafo resultante será ≤ que la original (porque existe la posibilidad de que un camino de dos pasos pesados sea menor que el camino directo). Para entender la relación entre el grafo original y su versión *dijksterizada* pongamos un ejemplo sencillo. En la Figura 2.5 se presenta un grafo circular simple con aristas dirigidas; cada nodo está numerado del 1 al 6. Como se puede viajar solo en sentido antihorario, el camino mínimo para ir de 1 a 6 es pasando por todos los demás mientras que de 6 a 1 se puede viajar directo. En la misma figura se incluye la representación matricial del grafo antes y después de aplicar el algoritmo de Dijkstra. Resulta intuitivo que si el grafo original es conexo, entonces su versión dijksterizada será completa. Tendremos, de esta manera, definida una distancia para todo par de palabras (*i*,*j*). Cabe notar que el método es sensible a qué palabras se incluyan en la lista que define la matriz. La inclusión o no de una palabra que esté altamente conectada (*hub*) puede alterar apreciablemente la estructura de la matriz *dijksterizada*.

Resumiendo, los pasos tipo receta de Doña Petrona para armar la matriz de distancia son los siguientes. Busque la mayor cantidad de textos que pueda, dentro de lo posible bien jugosos. Límpielos con cuidado. Arme la lista de palabras para la cual quiere obtener las distancias, a gusto. Cuente cuantas veces aparece cada par de palabras juntas y cuantas separadas. Revuelva a fuego lento usando

la ecuación **(2.4)**. Deje reposar y agregue un poco de computador holandés para eliminar el gusto amargo de los infinitos, y bualá.

Ya estamos en condiciones de estrenar nuestra nueva métrica. En el próximo capítulo describiremos algunos experimentos de asociación libre llevados acabo y analizaremos las secuencias producidas bajo la lupa de la matriz de distancias.

Capítulo 3 - Difusión en el espacio de palabras

En el presente capítulo describiremos dos versiones de un juego-experimento en el que los participantes generan secuencias de palabra mediante asociaciones libres. Analizaremos dichas secuencias como trayectorias en el grafo recién definido al tiempo que estaremos evaluando la utilidad de la métrica como herramienta de análisis.

3.1- Configuración Experimental: versión papel y versión Web.

La mecánica de estos experimentos difiere poco de la original propuesta por Jung a principios de siglo. Los sujetos reciben ciertas palabras como estimulo y se les pide que piensen la primera palabra distinta que se les ocurra. La diferencia fundamental con los experimentos de Jung es que la palabra asociada por un sujeto es usada como estimulo para otro, lo que le confiere al experimento cierta dinámica de juego.

1.1.2 - Versión Papel.

La primera implementación de este experimento, fue llevada a cabo durante el transcurso del IX Taller de Neurociencias en su versión en papel. Los sujetos estaban divididos en mesas de a cuatro jugadores; cada uno recibía palabras de su izquierda y enviaba las palabras asociadas hacia el jugador de la derecha. Las palabras asociadas eran registradas en las libretas de juego (una por jugador) de las cuales había, a su vez, dos versiones: una para generar tres palabras por cada palabra recibida y otra para generar solo una. En la Figura 3.1 pueden verse los dos tipos de libreta de juego. Los sujetos eran instruidos para que escribieran el primer sustantivo que se les ocurriera partiendo de la palabra que recibieran del jugador a su izquierda. Cada jugador recibía al comienzo una palabra raíz para iniciar el juego. Un experimento consistía de doce rondas generando un total de 48 palabras (12 palabras por trayectoria, una iniciada por cada sujeto).

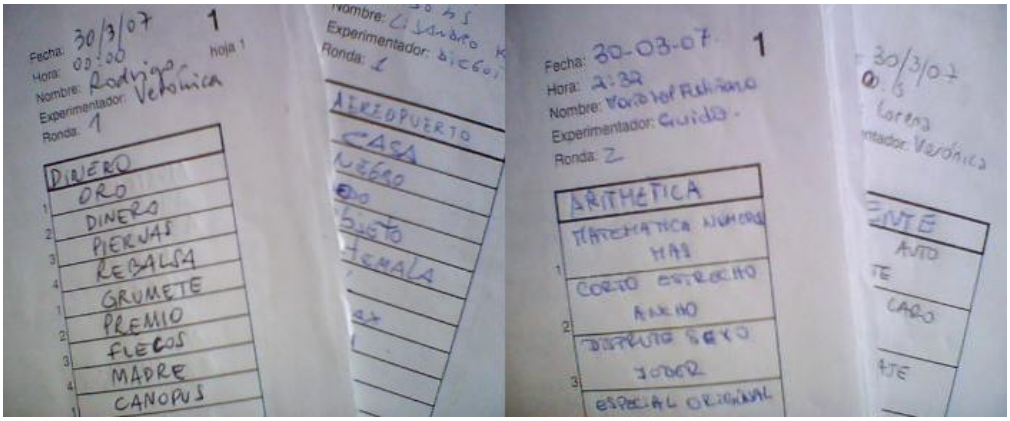

Figura 3.1: Libretas de Juego para registrar las palabras asociadas durante los experimentos de asociación libre realizados en el IX Neurotaller.

3.1.1 - Versión Web

Esencialmente la idea del experimento es la misma pero se agiliza la recolección de datos y su posterior procesado. El funcionamiento es el siguiente. Los sujetos que deseen participar se inscriben en: http://www.neurociencia.df.uba.ar/exp_pals/V2/registro.html completando simplemente nombre y dirección de correo. Luego de la inscripción reciben un mail de bienvenida explicando las reglas del juego así como también un hipervínculo a la pantalla de juego. El contenido del mismo puede verse en el anexo B. Al ingresar por primera vez, ven solo una palabra que será la raíz de un árbol de trayectorias. Normalmente encuentran allí todas las palabras que hayan recibido de otros jugadores hasta el momento. Una vez que

eligen alguno de los jugadores que les han enviado palabras pasan a otra pantalla en la que ven únicamente la palabra recibida y una cantidad de casillas para escribir que puede ser una o tres. Allí, deben escribir la(s) primera(s) palabra(s) que asocien con la palabra estímulo recibida. Luego seleccionan al jugador al cual se enviará la palabra asociada. Se la da la posibilidad a un jugador cada tanto, con una cierta probabilidad que puede ser controlada, de que envíe su palabra a dos jugadores distintos. Esto produce el desdoblamiento de una trayectoria en dos; por eso, cada palabra raíz genera no una sino un árbol de trayectorias. De esta manera es posible regular el crecimiento del número de trayectorias en juego. En la Figura 3.5 se presenta la secuencia de envío de palabras a partir de la pantalla de juego.

El desarrollo del juego puede ser monitoreado desde un sitio Web *backdoor* que permite: ver cuantas palabras pendientes tiene cada jugador, cuantas han sido enviadas día a día, enviar recordatorios a las casillas de correo de los jugadores avisando el número de palabras pendientes y enviar mails de penalización por no contestar palabras para evitar que se estanque el juego. Una vista en miniatura del sitio de control puede verse en la Figura 3.2. Todo el funcionamiento del juego está programado usando una combinación de HTML, PHP y JavaScript. Los datos son guardados en una base de datos usando un servidor estándar de MySQL.

Finalmente es posible extraer los datos, e indexar cada palabra para conformar la matriz de distancias como se describe en el apartado 2.3. En la Figura 3.5 se presentan, a modo de muestra, las primeras palabras de un árbol generado en el experimento. A simple vista, resulta intuitivo que hay estructura detrás de dichas secuencias; por ejemplo repeticiones como TREN, VIA, TREN o el la percepción de que nunca se abandona un vecindario de palabras (se vuelve a AUTO después de muchos pasos). A continuación nos dedicaremos a analizar estas secuencias utilizando como herramienta la métrica

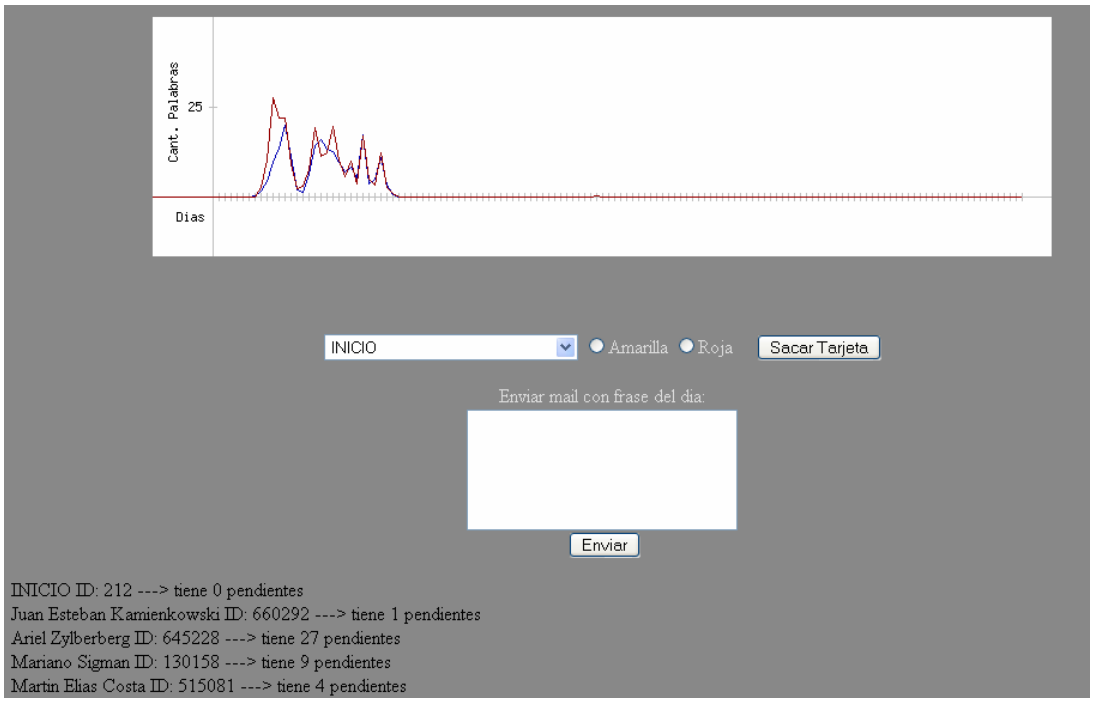

Figura 3.2: Vista en miniatura del sitio *backdoor* de monitoreo.

definida en el Capítulo 2. El objetivo será doble: ver qué nos es posible inferir, a partir de las secuencias, sobre el proceso de transición espontánea entre representaciones verbales así como también evaluar la noción de distancia que hemos propuesto como herramienta de análisis.

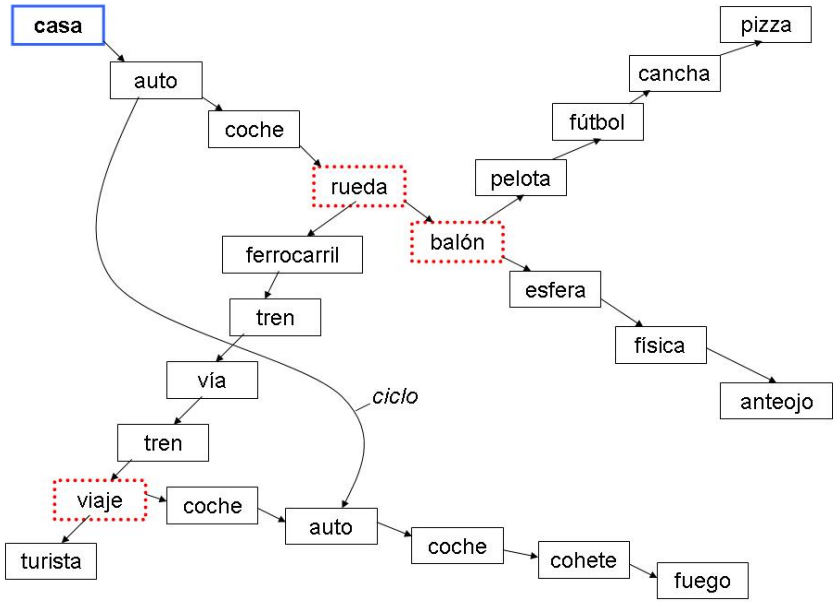

Figura 3.3: Árbol de trayectorias generado en el experimento de asociaciones. En línea de puntos se marcan las bifurcaciones. A un corte de una trayectoria con ella misma lo llamaremos ciclo.

Ambos juegos fueron jugados en una segunda variante, que tenia como objetivo estudiar secuencias de palabras generadas por distintos mecanismos. En esta segunda variante cada jugador recibía, tal como en el juego original, una palabra "semilla". El jugador tenia que pensar tres palabras asociados con la semilla, escribirlas y pasar la ultima al jugador siguiente. Esta variación inocente debería generar predicciones muy distintas en la estructura de las trayectorias. Intuitivamente, las tres palabras asociadas

deberían formar una vecindad que rodee a la semilla, en vez de una trayectoria que "difunde" desde la semilla. El tiempo no nos ha permitido completar el análisis de este segundo experimento, aquí meramente incluimos un ejemplo de dicha trayectoria que captura, creemos, la idea que queríamos explorar (ver Figura 3.4).

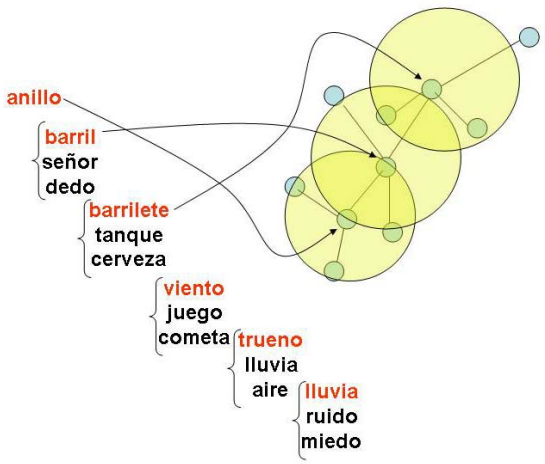

Figura 3.4: Ejemplo de una secuencia generada en la segunda variante del experimento (pensar de a tres palabras)

.

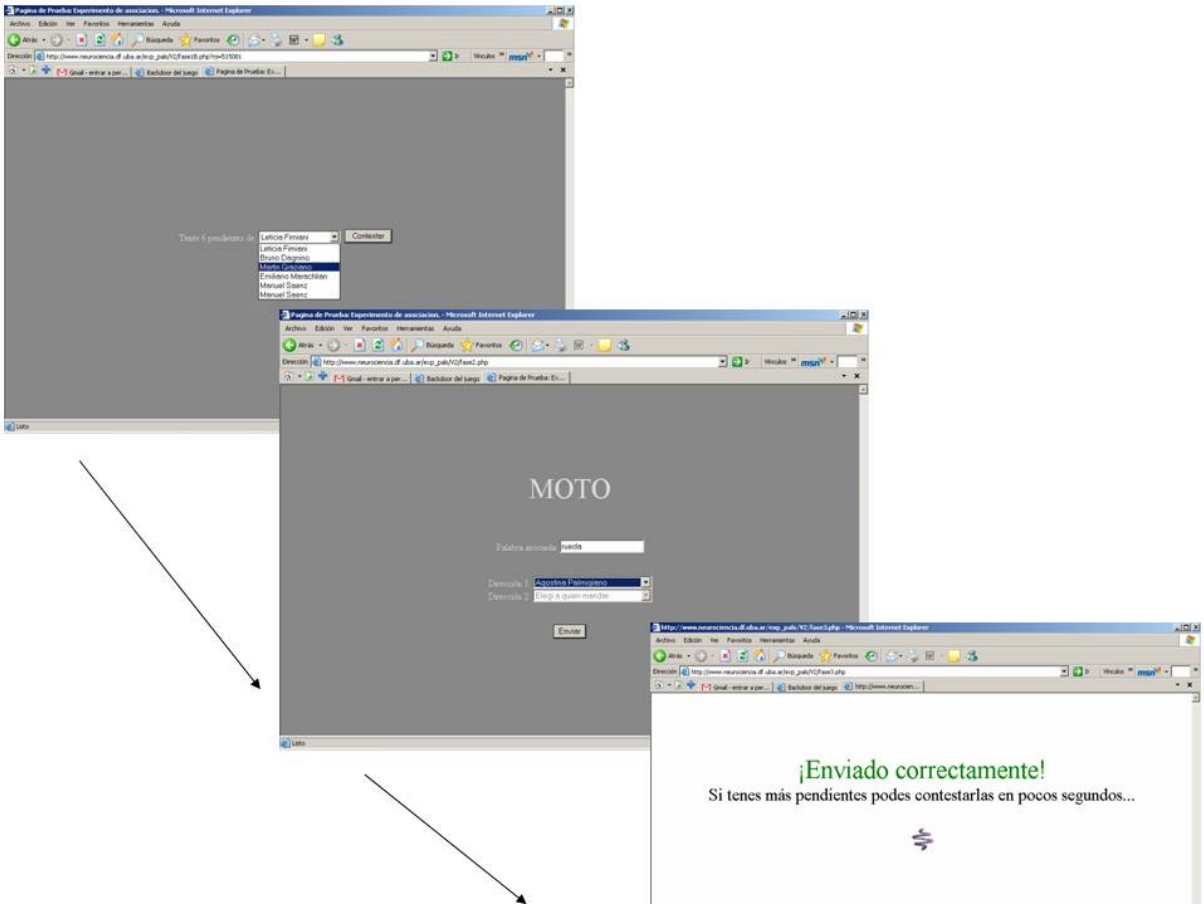

Figura 3.5: Secuencia de envío de una palabra a partir de la pantalla de juego.

3.2- Análisis de resultados

Para ganar intuición sobre las diferencias y similitudes que hay entre la estadística de palabras en texto y las producidas por asociación libre, hagamos algunas observaciones cualitativas. En la Figura 3.6 (columna izquierda) se muestra el podio de las diez palabras que más aparecieron durante el juego. La columna de la derecha tiene, de todas las palabras del juego, la diez más frecuentes **para el corpus.** Esta diferencia es esperable y resulta evidente si consideramos, por ejemplo, la preposición "a". Es una palabra

extremada mente frecuente tanto en el habla como en texto escrito pero no aparece nunca en un experimento de asociación. Por otra parte, esta tabla también pone en evidencia una dificultad intrínseca con la que hemos lidiado en este experimento. La palabra "te", aparecida en el juego, probablemente se refería a la infusión. La alta frecuencia de esta palabra evidentemente se refiere al pronombre personal "te". En algunas ocasiones, los acentos distinguen dos significados de la misma palabra, pero en este trabajo, dado que numerosas fuentes con las que contábamos no contaban con buena acentuación, decidimos considerar todas las palabras que se distinguen por un acento como la misma ocurrencia.

Juego (web)	Corpus
'cama'	'te'
'musica'	'bien'
'arbol'	'hasta'
'sol'	'don'
'perro'	'casa'
'agua'	'dia'
'juego'	'hombre'
'cuchillo'	'tiempo'
'calor'	'nada'
'noche'	'dios'

Juego (taller)	Corpus
'sexo'	'dos'
'negro'	'bien'
'blanco'	'casa'
'mujer'	'dia'
'vino'	'otro'
'muerte'	'hombre'
'agua'	'tiempo'
'mar'	'nada'
'azul'	'dios'
'alcohol'	'vida'

Figura 3.6: Podio de las palabras que aparecieron durante el juego: para el juego y para el corpus

1.1.3- Desplazamiento medio

Si esperamos que la métrica definida en el Capítulo 2 capture alguna característica del espacio de asociaciones, entonces, debería ocurrir que la distancia entre primeros saltos de una trayectoria sea pequeña. Pero, ¿pequeña comparada con que? Se la puede comparar con cualquier otro salto posible entre palabras que hayan aparecido en el juego. En la Figura 3.7 se presentan el histograma de primeros saltos de todas las trayectorias así como también la distribución de distancias en la matriz *dijksterizada* (ver sección 2.4). Efectivamente, lo que se observa es que se han elegido conectar palabras que están más cerca que la distancia media en el grafo. Asumiendo la hipótesis de normalidad, las medias estimadas para la

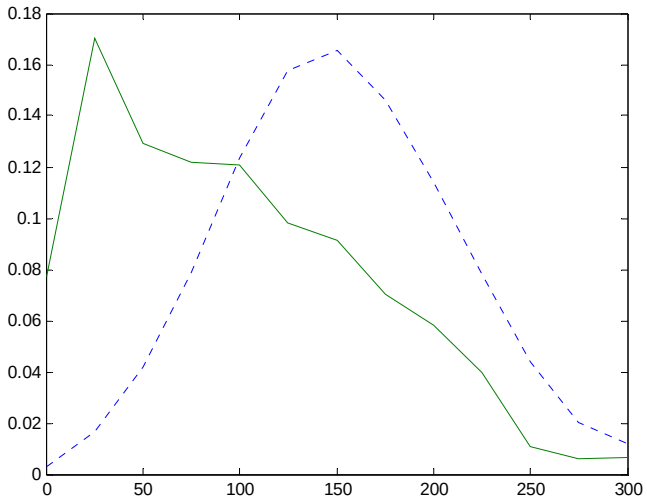

Figura 3.7: Histogramas de distancias de saltos a primeros vecinos (continua) y de distancias medias en la matriz de la métrica (punteada)

distribución de primeros saltos y de la matriz de distancias son: 97 ± 2 y 151 ± 1, respectivamente.

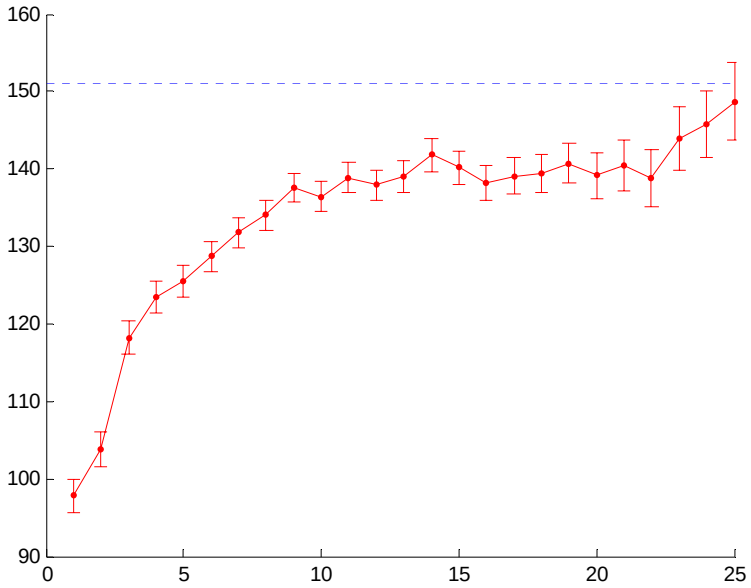

Figura 3.8: Distancia media entre saltos en función del orden del salto. En línea de puntos, la media del grafo.

Así como esperábamos que los saltos a primeros vecinos fueran menores que la media del grafo, la distancia luego de saltar dos veces también debería ser menor que la media

pero mayor que la distancia de los primeros saltos. En general, esperamos que se vaya perdiendo la correlación entre palabras a medida que se toman saltos de orden mayor. Esto podemos visualizarlo si graficamos la media de la distribución de saltos en función del orden del salto. Tal es el gráfico que se muestra en la

Figura 3.8. Se ve que luego de aproximadamente 10 saltos las palabras están, esencialmente, decorrelacionadas; o en términos lingüísticos, pertenecen a contextos diferentes. Una imagen mental posible para entender el comportamiento de la curva es pensar en una partícula que difunde y se aleja lentamente de su vecindario pero en un espacio acotado, pues converge a la distancia media en el grafo.

Si bien el gráfico de la Figura 3.8 nos permite poner de manifiesto algunas de las características difusivas de las trayectorias e ir ganando confianza en nuestra definición de métrica, oculta un fenómeno interesante por el simple hecho de haber colapsado todos los histogramas de saltos en su media. La Figura 3.9 es, esencialmente la misma que la Figura 3.8 pero mostrando en código de colores todo el histograma para cada orden de salto. Notar lo que ocurre para saltos pares, hay una gran población de saltos a distancias cercanas a cero.

En realidad, mirando con detalle las poblaciones de saltos se ve que hay cantidades considerables que son exactamente cero. Ahora bien, la única forma de obtener un cero absoluto en nuestra métrica es volviendo a la misma palabra. Es decir que lo que estamos viendo son los cortes de la trayectoria con ella misma. Llamaremos un ciclo a la trayectoria cerrada que se forma entre cortes. Dado la frecuencia con que aparecen y la información que pueden proveernos de la organización del espacio de conceptos, nos

dedicaremos a estudiarlos en detalle en la sección siguiente. Para darnos una idea de lo que está ocurriendo podemos buscar cuales

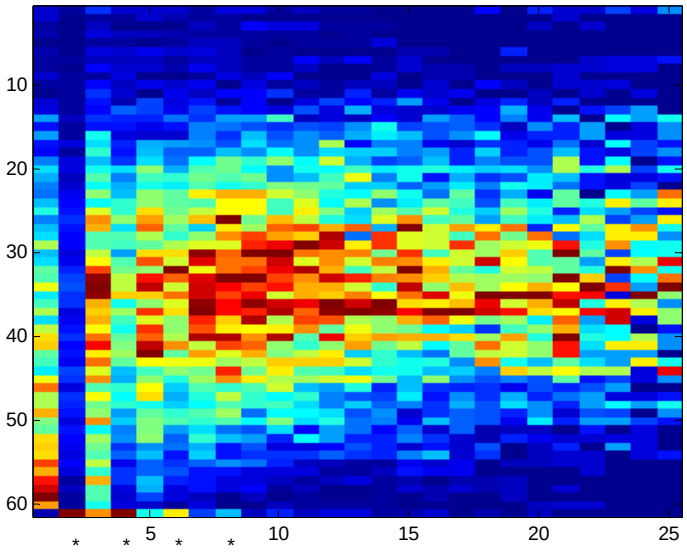

Figura 3.9: Histogramas de saltos en códigos de color en función del orden del salto. En los saltos marcados con un * se observan grandes poblaciones en distancias cercanas a cero.

son las palabras que ciclan. Esto es lo que se muestra en Figura 3.10. Se ve que la mayoría corresponde a pares de palabras en relación dicotómica. Sin embargo, existen también tríos de relaciones en los que se puede ir alternando en una secuencia tipo A-B-C-B-C-B-A-B-C-B…

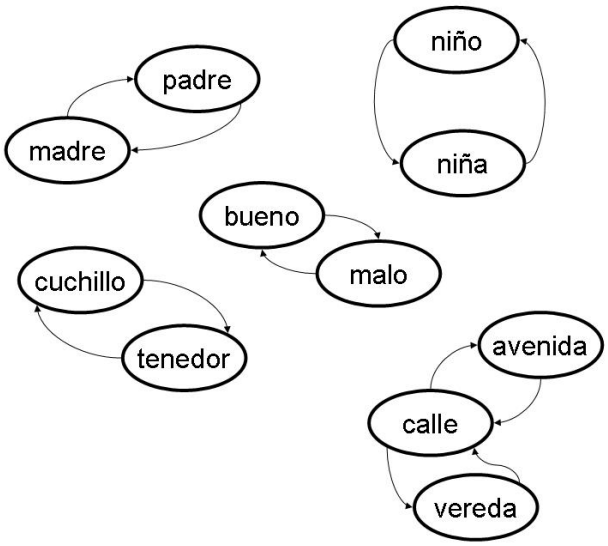

Figura 3.10: Palabras que aparecen frecuentemente formando ciclos.

3.2.1- Ciclos

En la sección anterior vimos que hay considerablemente más ciclos en los saltos pares que en los impares. Podemos hacer un gráfico de la fracción de ciclos en función del número de saltos (ver Figura 3.11). Como es de esperar no hay ningún ciclo de orden uno pues, como regla, estaba prohibido responder la misma palabra que se recibía. Puede verse que aproximadamente el 16% de las veces se regresa a la misma palabra luego de dos saltos. La forma del

histograma es bastante llamativa. Tiene un decaimiento, esperable, que indica que es poco probable volver a la palabra original luego de muchos saltos lo cual es consistente con la idea del modelo difusivo. Se observan, también, una serie de picos montados sobre ese decaimiento que aparecen en los números de salto par. Nuevamente, la representación del proceso como una caminata al azar puede ayudarnos a comprender algo sobre la organización de los nodos. En la Figura 3.11 muestra la estructura de ciclos para una red bidimensional cuadrada. El número de ciclos de orden dos es igual a la inversa del número de primeros vecinos.

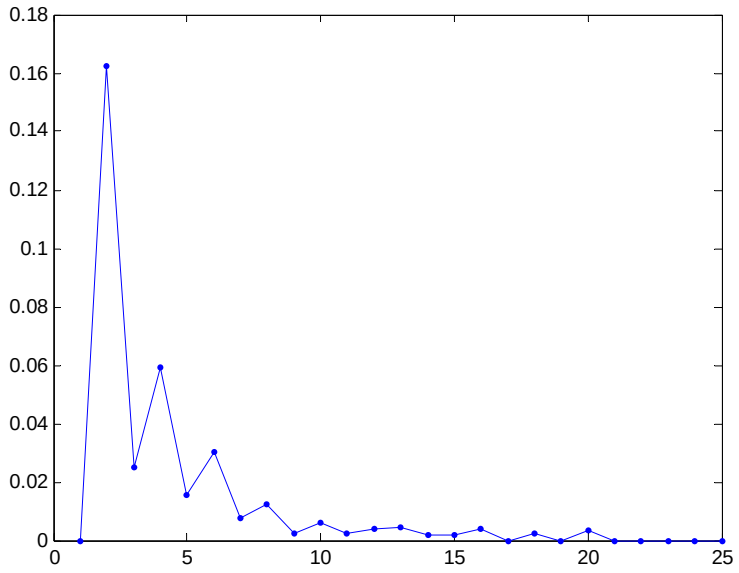

Figura 3.11: Histograma de ciclos en función del orden del salto.

Dada la estructura de la red es imposible volver al mismo punto luego de un número impar de saltos. Podemos, entonces, tratar de agregar más vecinos con la esperanza de cambiar el aspecto de la curva. Si lo hacemos, convirtiendo a la red cuadrada en una cúbica, siguen estando prohibidos los ciclos impares. Una forma de tener ciclos impares es agregar los vecinos de las diagonales de los cuadrados. En la Figura 3.13 (b) se puede ver la estructura de ciclos para dicha red. La diferencia entre las dos formas de agregar vecinos está relacionado con la siguiente pregunta: ¿los amigos de mis amigos, son también amigos míos? En el caso de la red cuadrada la respuesta es no, mientras que para la red con las diagonales podemos responder que algunos sí. Una forma de cuantificar esto es usar el coeficiente de clustering que mide el porcentaje de links que hay en un grupo de vecinos de un nodo respecto del grafo completo que se puede formar con esos mismos nodos. Esa es la forma de definirlo para grafos discretos, existen distintas propuestas sobre como generalizar este coeficientes para grafos pesados (en el trabajo de Kalna y Higham, 2006, se discuten distintas alternativas). Una manera intuitiva de hacerla es discretizando el grafo tomando solo un número de primeros vecinos. En la Figura 3.12 se muestra como cambia el coeficiente de clustering para distintos cortes de primeros vecinos. Se separa también la media de sobre todos los nodos (en continua) y la media

sobre las palabras que fueron raíces (en punteado) del experimento. Esto pone en evidencia el hecho de que el grafo es fuertemente no homogéneo. Como contraste, las redes bidimensionales que analizamos a modo de ejemplo, sí son homogéneas; es decir que da lo mismo empezar una caminata al azar de cualquiera de sus nodos. En el caso de la matriz de distancias esto no es así y vamos a tener que tenerlo en cuenta a la hora de realizar simulaciones.

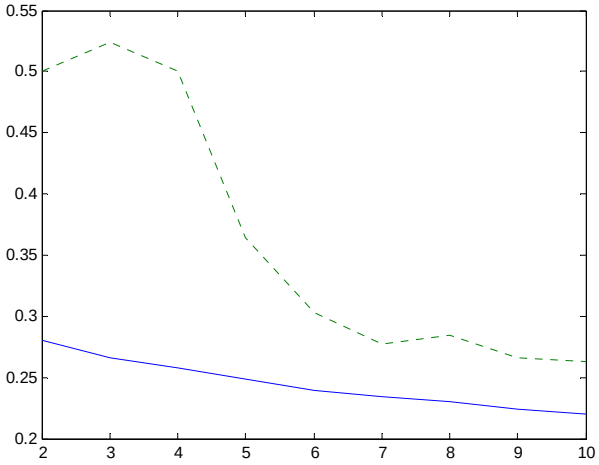

Figura 3.12: Coeficiente de clustering en función del número de primeros vecinos considerados. En línea continua, la media para todo el grafo y en punteada solo para las palabras que fueron raíces del experimento.

A lo largo de este capítulo hemos ido analizado algunas características de las trayectorias, y repetidas veces vimos que resulto útil interpretar esas trayectorias como un proceso difusivo en el grafo semántico. En el capítulo siguiente intentaremos reproducir algunas de estas características modelando las trayectorias como paseos al azar en el espacio de palabras.

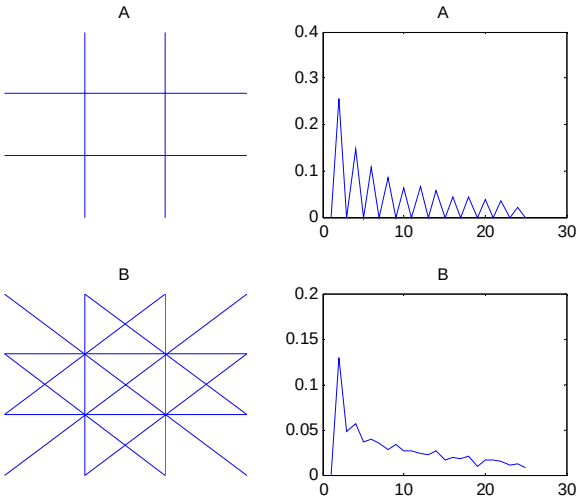

Figura 3.13: Simulación de la estructura de ciclos en dos tipos de redes: (a) red cuadrada bidimensional, (b) misma red agregando los vértices diagonales.

Capítulo 4 Paseando al azar en una red semántica

En el capítulo anterior fuimos ganando confianza sobre la imagen de las trayectorias como procesos difusivos en el grafo semántico. Siguiendo esta idea, intentaremos reproducir algunas de las características de las trayectorias analizadas en el Capítulo 3 simulando paseos al azar en la matriz de distancias. Existe la arbitrariedad de elegir las probabilidades de salto a partir de los elementos de matriz Dij. Analizaremos tres asignaciones distintas. En todos los modelos tendremos un parámetro libre que determina la estocasticidad del proceso. Finalmente ajustaremos las simulaciones a los datos utilizando este parámetro.

4.1- Modelos alternativos

4.1.1- Modelo de probabilidades exponenciales

En este modelo asignamos a cada vecino una probabilidad de salto proporcional al factor de Boltzmann, es decir:

$$P_{ij} \propto e^{-B.D_{ij}} \qquad \textbf{(4.1)}$$

Dada esta asignación, el salto más probable es al vecino más cercano. Para B tendiendo a infinito la distribución se vuelve completamente determinista y siempre se viaja al vecino más cercano. Por el contrario cuando B tiende a cero, la distribución es uniforme y se puede saltar completamente al azar entre cualquier par de vecinos.

4.1.2- Modelo de umbral de vecinos

Aquí proponemos una distribución de probabilidades uniforme para los primeros K vecinos más cercanos. Nuevamente existen dos límites. Cuando K vale uno, solo es posible saltar al vecino más cercano mientras que si K vale N (el número total de nodos) recuperamos la distribución uniforme para cualquier par de vecinos.

4.1.3- Modelo de nivel de ruido variable.

Finalmente, proponemos un tercer modelo en el cual no definimos las probabilidades a priori, sino que elegimos el siguiente paso de la simulación de manera operativa (tipo Montecarlo). Utilizando un generador de números pseudoaleatorios con distribución uniforme en el intervalo [0,1] escogemos al próximo vecino con el siguiente criterio:

$$siguiente(j) = Max\{\frac{B}{D_{ij}} + rand[0,1]\} \qquad \textbf{(4.2)}$$

Los límites B → 0 y B → ∞ coinciden con los del modelo exponencial. Este modelo también tiene un número de vecinos bien definidos mediante:

$$\frac{B}{D_{\min(j)}} - \frac{B}{D_{ij}} < 1 \qquad \textbf{(4.3)}$$

Sin embargo, a diferencia del modelo 2, no todos los vecinos son equiprobables. El nombre elegido puede resultar un poco oscuro. La Figura 4.1 aclara un poco el asunto. En ella se muestra la inversa de la distancia por el factor B (en este caso B= 6) para algunos nodos ordenada de menor a mayor. Las líneas de puntos muestran el límite de la condición **(4.3)**. Solo los nodos que queden a la izquierda de la línea horizontal pueden llegar a confundirse con el máximo al sumar el "ruido" aleatorio.

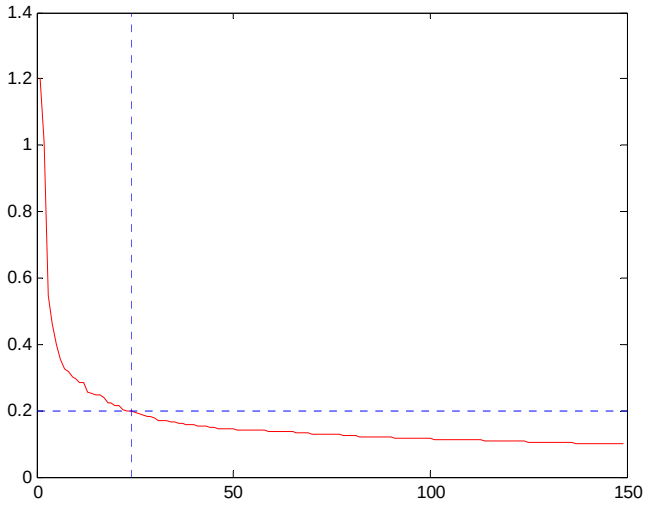

Figura 4.1: Inversa de las distancias ordenadas sin ruido. Las líneas de puntos marcan los límites de la condición (4.3).

4.2- Simulaciones y ajuste a los datos

La forma de realizar un paseo al azar es la siguiente. Elegimos un nodo del que partiremos (digamos *j*) y ponemos en una bolsa papelitos con los números de los nodos a los que se puede viajar desde *j*. Tenemos que poner en la bolsa los papelitos en proporciones tales que cuando saquemos uno (digamos *i*) lo hagamos con las probabilidades que requiere cada modelo. Para dar el siguiente paso es necesario preparar una nueva bolsa, una que siga las probabilidades de partir del nodo *i*. La imagen de la bolsa nos ayuda a entender la idea detrás de la caminata al azar pero no resulta una alternativa conveniente (consume mucho tiempo ¡y papel!) Por suerte, partiendo de un generador de números al azar

con distribución uniforme en el intervalo [0,1] es posible obtener números aleatorios con una distribución arbitraria.

Una cosa que es importante notar es que dada la inhomogeneidad del espacio no da lo mismo que tomemos cualquier nodo para empezar una trayectoria. Si esperamos reproducir algunos de los comportamientos que observamos en el Capítulo 3 será necesario empezar las caminatas desde las mismas raíces que usamos en los experimentos.

Cuando describimos los modelos mencionamos que los tres poseen un limite completamente deterministico (en el cual siempre se viaja a la palabra más cercana) y un limite completamente aleatorio (que corresponde a sacar palabras con probabilidad 1/N). Podemos comprobar esto simulando caminatas en estos límites. En la Figura 4.2 se muestran las curvas de difusión y ciclos para los tres modelos en el limite deterministico (las tres se superponen). Por otro lado, en la Figura 4.3 se presentan los límites aleatorios. La línea de puntos corresponde a la media de las distancias del grafo. Se puede ver que los tres dan resultados equivalentes. La probabilidad de obtener un ciclo es tan baja que nada se observa en el grafico de la derecha. La pregunta que sigue, lógicamente, es si para valores intermedios de los parámetros que gobiernan la estocasticidad obtendremos curvas de difusión y ciclos similares a los experimentos.

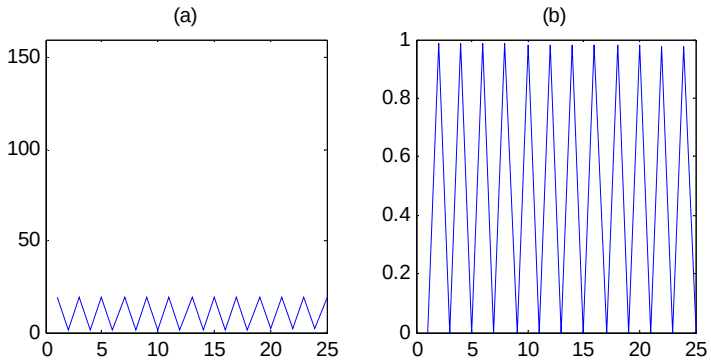

Figura 4.2: Gráficos de difusión (a) y de ciclos (b) para los tres modelos en el límite determinista (las curvas se superponen)

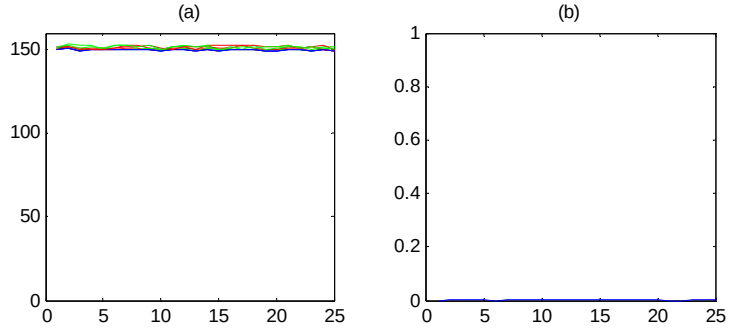

Figura 4.3: Gráficos de difusión (a) y de ciclos (b) para los tres modelos en el límite aleatorio. En línea de puntos se grafica la media de la matriz de distancias. Se puede ver que los tres modelos dan resultados equivalentes. La probabilidad de obtener un ciclo en este límite es bajísima.

Para responder a esta pregunta, podemos encontrar el parámetro óptimo para cada uno de ellos buscando el que minimiza la distancia a las curvas de difusión y ciclos conjuntamente (en el sentido de cuadrados mínimos). En la Figura 4.4 se presentan las curvas obtenidas para los parámetros óptimos (en azul) junto con los datos (en rojo y con puntos) para los tres modelos. Se puede ver que los tres logran reproducir medianamente bien la curva de

difusión, no así la curva de ciclos. Debido a la estructura de mundo pequeño del grafo es esperable que cualquier regla que asigne mayor probabilidad a los vecinos más cercanos dé como resultado una curva de difusión razonable. Esto se debe a que en pocos pasos se llega al borde del espacio y distintas formas funcionales se vuelven rápidamente indistinguibles. La estructura de ciclos aparece, entonces, como una herramienta para distinguir entre ellos. El modelo exponencial (2) no parece reproducir el comportamiento de los ciclos ni para saltos de orden bajos ni en la cola del histograma. Tampoco presenta la estructura de serrucho para los saltos de orden par.

Por otro lado, si observamos las curvas de los otros dos, vemos que el de ruido variable parece reproducir bien el comportamiento de los ciclos de orden bajo mientras que el modelo 2 se ajusta adecuadamente a la cola del histograma.

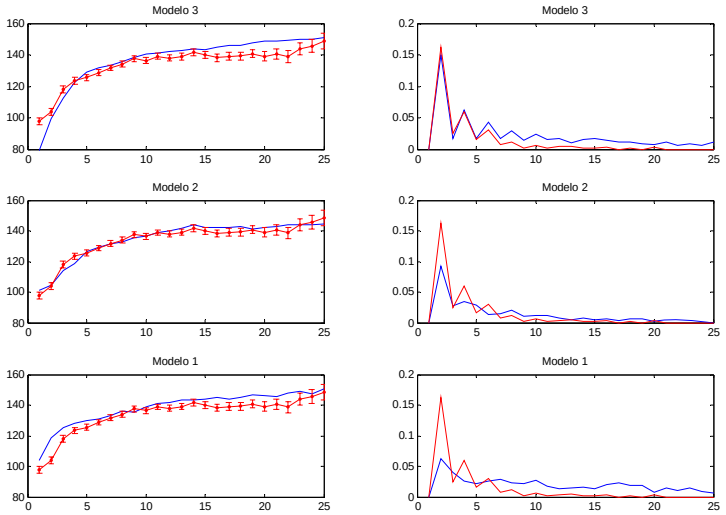

Figura 4.4: Ajustes de los modelos a los datos. (Izq) Curvas de difusión (Der) Histogramas de ciclos.

Supongamos que el hecho de que el modelo 3 explique bien los saltos de orden bajos esta relacionado con el hecho de que da más peso (y sobre todo un peso mas uniforme) a los vecinos cercanos. Por otro lado, la cola del histograma queda mejor explicada por el modelo que asigna pesos uniformes a los vecinos y por lo tanto permite cada tanto saltos de mayor longitud. Estaríamos tentados de combinar los dos modelos, por ejemplo, permitiéndole al modelo 3, cada tanto, realizar un salto con probabilidad uniforme. En la Figura 4.5 se muestra el histograma de ciclos para un modelo de ese tipo, la curva de difusión no cambia apreciablemente. Se ve que puede reproducirse el comportamiento de la curva en todo el rango de saltos, pero esto lo hemos logrado al costo de incluir un nuevo

parámetro a ajustar y dado que tanto la curva que da solo el modelo 2 como la que da solo el 3 son razonables no parece valer la pena. Lo interesante detrás de esto es que sugiere que el proceso de asociación es una especie de caminata al azar con saltos. Uno difunde localmente y cada tanto una conexión larga nos manda a otro lugar del mundo donde volvemos a difundir localmente. Una posible herramienta formal para modelar este proceso, definida por el matemático francés Paúl Pierre Levy, son los "levy flights" que describen una caminata al azar donde la distancia de la probabilidad de los saltos esta dada por una función con "cola larga". Este formalismo ha sido utilizado para describir numerosos procesos de transporte, entre ellos el transporte humano a lo largo del planeta. Brockman et al. (2006) midieron la difusión de personas rastreando el camino de billetes a través de sus números de serie. Este fenómeno que combina los más frecuentes viajes a pie, o en bicicleta o en coche, con esporádicos viajes en avión que nos cruzan de un lado al otro del mundo queda bien descripto por la matemática de Levy. En la Figura 4.6 se puede ver la diferencia entre un proceso difusivo Browniano y un proceso de Levy. Notar los saltos de largo alcance que ocurren cada tanto.

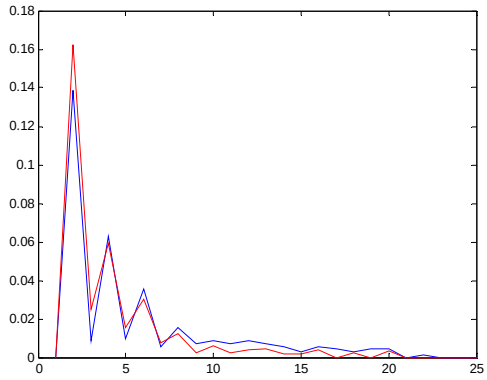

Figura 4.5: Histograma de ciclos para el modelo difusivo con saltos; datos en rojo.

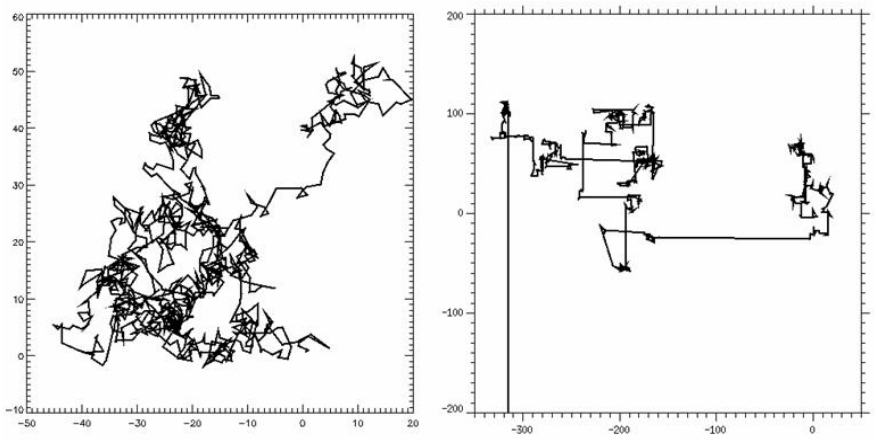

Figura 4.6: Random walk clásico Browniano (izq) y de Lèvy (der)

En un futuro, nos proponemos poder colapsar el análisis aquí descripto a un modelo formal capaz de capturar los aspectos críticos de la dinámica de las secuencias de asociaciones libres de palabras. Si bien, el análisis desarrollado hasta aquí no permite asegurar

cuales son los detalles del mecanismo de asociación libre si podemos concluir que es posible entenderlo y modelarlo como un proceso difusivo. Al mismo tiempo, hemos mostrado que el grado de pertenencia de dos palabras a un dado contexto es pasible de ser cuantificado. Todo esto indica una dirección a seguir y motiva refinar los experimentos y aumentar el tamaño del corpus para mejorar la convergencia del conteo.

A modo de cierre y como última figura del trabajo mostramos una trayectoria generada por el modelo difusivo con saltos.

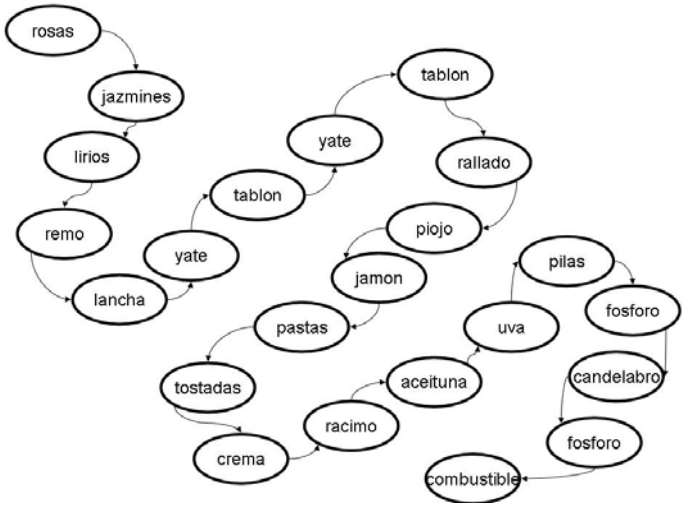

Figura 4.7: Trayectoria generada con el modelo difusivo con saltos.

Capítulo 5 - Resumen, Conclusiones y algunas consideraciones hacia el futuro

En el presenta trabajo hemos intentado cuantificar el grado de pertenencia de dos palabras al mismo contexto; específicamente definimos una métrica entre palabras a partir de la estadística de textos. Utilizamos la misma como herramienta de análisis de secuencias de palabras generadas por asociación libre para intentar estudiar algunas características de la organización del léxico.

Lo técnico

En el transcurso aprendimos ciertas consideraciones técnicas importantes y generamos algoritmos que pueden ser utilizables para futuros proyectos relacionadosy que detallamos a continuación:

- Lecciones sobre que corpus usar y cuales no. Las bases de datos controladas por empresas (como por ejemplo Google) utilizan algoritmos no transparentes, con mecanismos de redondeo, con

criterios de prioridades que hacen que estas no puedan ser utilizadas como base de una métrica sólida. Es mejor usar texto directamente.

-Desarrollo de la estructura de datos: creamos una plataforma flexible para generar una amplia (y creciente) base de datos sobre distintos corpus: Generamos un software (en PERL) capaz de bajar textos (de done sea), clasificarlos (según si son escritos, periodísticos, orales etc…) limpiarlos y contar co-ocurrencias separadas por un numero variable de palabras. Un aspecto importante de esta estructura es que permite incorporar nuevos elementos para generar estadística sobre un corpus creciente. Un aspecto para el futuro será consolidar esta base de datos cuya convergencia aun no hemos caracterizado con precisión. Por otra parte, otro elemento para el futuro será estudiar la sensibilidad de la métricas al uso de corpus de distinto tipo (texto Web, discurso oral, poesía, narrativa, periodístico, época del texto…) Usamos el algoritmo de Dijkstra como un operador de grafos. Esto permite trabajar con un grafo en el cual la distancia entre todo par de nodos es finita.

- Desarrollo de una plataforma experimental: Creamos una interfaz Web sencilla que permite generar experimentos online variables. De esta manera es muy fácil conseguir sujetos para realizar experimentos, pues lo hacen desde su casa y el volumen de datos que podemos conseguir es ilimitado. La misma será utilizada

en un futuro para realizar variantes de los experimentos aquí descriptos; más adelante discutimos algunas de ellas.

Lo que le interesaría a Jung

Partiendo de sus ideas usamos experimentos de asociación para explorar el espacio de palabras, pero de manera cuantitativa. ¿Qué aprendimos sobre las palabras y su organización en nuestras cabezas? A continuación resumimos algunos de los resultados:

-Midiendo estructura donde todos la ven: Pudimos medir y cuantificar un hecho evidente para todos que es que las secuencias de palabras generadas por asociación espontánea tienen estructura.

-Similitudes y diferencias entre distintos mecanismos de generación de palabras.

Pudimos medir y cuantificar un hecho casi evidente para todos que es que esta estructura está en relación con el orden de ocurrencias de palabras en el lenguaje escrito. Si bien esto indica la correlación entre estas dos manifestaciones cognitivas del lenguaje (si se puede llamar lenguaje a las secuencias de palabras generadas por asociación libre) pudimos observar algunas diferencias importantes, empezando por la distribución de palabras mas frecuentes. Esto indica que, aun en el mejor de los esfuerzos para

explicar los datos comportamentales como un mecanismo de transporte dentro del grafo de métrica de co-ocurrencia nunca podremos explicar la totalidad de la estructura de las secuencias a partir de nuestra métrica (se escapan, por ejemplo, asociaciones fonológicas; uno no usa la misma estadística cuando habla que cuando escribe, etc.)

-Trazas de un mecanismo de trasporte difusivo y ergódico.

Identificamos variables robustas que caracterizan la estructura de estas secuencias y que caracterizan algunos aspectos del mecanismo de transporte. A saber:

-Parece un proceso difusivo de algún tipo. La distancia entre pares de palabras aumenta montónicamente con su separación en la secuencia.

-El numero de elementos para llegar a un valor asintótico es 10 (no es ni 3 ni 100), lo cual determina el orden de magnitud.

-Llega a un valor asintótico que coincide con el valor medio del grafo. Esto indica que uno no converge hacia el centro del grafo, como podría haber sido (dado que estamos estudiando transitorios y por lo tanto si hubiese atracotes veríamos la convergencia al

atractor y no la permanencia ahí) y sugiere además que no hay atractores (podría ser que estos estén repartidos por el grafo separados aproximadamente por la distancia media lo cual no podríamos distinguir con este análisis)

-Caracterizamos un modelo de transporte que captura aspectos esenciales de los datos.

Hay un factor de escala critico que es como se ecualizan las probabilidades y encontramos que el modelo que mejor se ajusta a los datos corresponde a una caminata al azar con saltos. Una especie de versión casera de los “lèvy flight”. Sin embargo, esto no es concluyente y hay otros modelos que explican razonablemente bien los datos. Esto motiva a refinar los experimentos y aumentar el tamaño del corpus a fin de poder distinguir ellos.

El Futuro:

Proponemos extender el espectro de experimentos y estudiar distintos mecanismos generativos. Ya tenemos hecho el de generación de tres significados que nos permitirá estudiar si, como se puede intuir, cuando uno piensa palabras alrededor de una semilla estas establecen una “bola”.

Estos experimentos han motivado una colaboración con el INN en brasil, donde generamos las mismas listas en personas que se levantan del sueño REM. La idea (de Jung) es que distintos mecanismos de asociación puedan revelar distintos estados de conciencia. Pretendemos establecer una medida consistente y robusta e informativa sobre los mecanismos generadores de esta afirmación.

Llevaremos adelante también, en breve, experimentos de fisiología en los que esta métrica puede ser usada como un regresor para estudiar transiciones entre distintos eventos. En general los experimentos hechos en espacios de dimensionalidad alta son muy interesantes pero muy difíciles precisamente por que uno no cuenta con una buena métrica. La idea es simple, persona asociando, EEG registrando y entender la fisiología de transiciones frecuentes, infrecuentes, etc… o sea mapear las transiciones entre estados del EEG al grafo.

Referencias

Brockmann D., Hufnagel L., Geisel T., *The scaling laws of human travel*, Vol 439|26 January 2006|doi:10.1038/nature04292

Changizi M, *The global organization and economy of the lexicon*, en preparación.
Diario La Nacion, Notas entre 2001 y 2007.

Dijkstra E.W., *A note on two problems related with graphs*, Numerische Mathematik, Springer, 1959

Fellbaum C., *WordNet: An Electronic Lexical Database*, MIT Press, 1998, ISBN-10:0-262-06197-X

Ferrer i Cancho R., Solé R., *The small world of human language*, Proceedings of the Royal Society of London B, 2001

Gilfillian I., La biblia de MySQL, ANAYA, 2003, ISBN: 84-415-1558-1

Google, buscador en internet: http://www.google.com.ar

Gutemberg Project, http://www.gutemberg.org

Hopfield J.J., *Searching for memories, Sudoku, implicit check-bits, and the iterative use of not-always-correct rapid neural computation*, arXiv:q-bio/0609006v2 [q-bio.NC]

Jung C., *Gesammelte Werke: experimentelle untersuchungen* , ed. Walter-Verlag, 1991, ISBN-10: 353040702X.

Kalna G., Higham D.J:, *Clustering coefficients for weighted networks*, Glasgow: University of Strathclyde Mathematics Research, 2006

Lévy flight. (2007, October 13). In *Wikipedia, The Free Encyclopedia*, from http://en.wikipedia.org/w/index.php?title=L%C3%A9vy_flight&oldid=164271904

Meyer D.E., Schvaneveldt R.W., *Facilitation in recognizing words: Evidence of a dependence upon retrieval operations*, Journal of Experimental Psychology, 1971

NP-complete. (2008, March 13). In *Wikipedia, The Free Encyclopedia*, from http://en.wikipedia.org/w/index.php?title=NP-complete&oldid=197983133

Quine W., *Ontological Relativity and Other Essays,* Columbia Univ. Press, 1969, ISBN: 0-231-08357-2.

Saussure F., *Curso de lingüística general*, ed. Losada, 2001, ISBN: 950-03-6108-6

Shannon C.E. Weaver W. *The mathematical theory of communication*, University of Illinois Press, 1963

Sigman M., Cecchi G., *Global organization of the Wordnet lexicon*, Proc Natl Acad Sci U S A. 2002 February 5; 99(3): 1742–1747.

Zipf G.K., *Selected Studies of the Principle of Relative Frequency in Language*, Harvard University Press, 1932.

Anexo A ¿Por qué perl?

Perl es un lenguaje de propósito general originalmente desarrollado para la manipulación de texto y que ahora es utilizado para un amplio rango de tareas incluyendo administración de sistemas, desarrollo Web, programación en red, desarrollo de GUI y más. Es el lenguaje más utilizado para programar herramientas de extracción de información de textos (parsing). Debido a que posee una gran cantidad de funciones para la manipulación de cadenas así como una integración sencilla con la búsqueda de expresiones regulares es particularmente útil para la tarea de limpieza de los textos del corpus (ver - Norma de estadística en textos.). También permite un manejo simple de peticiones HTTP lo que resulta muy práctico a la hora de gestionar la bajada de textos de sitios Web como Gutemberg.

Anexo B: Mails del Juego

1) **Mail de invitación:**

Queremos invitarte a participar en un juego que todos los hermanos Pauls catalogaron como "...más adictivo que el Sudoku". El juego consiste en asociar palabras libremente y será jugado, en

una primera prueba, por integrantes de todos los claustros de la FECyN. Es, además, parte de un proyecto de investigación cuyos resultados serán divulgados una vez que el juego haya concluido. Para jugar basta una conexión a Internet y dos minutos al día son más que suficientes para no ver una tarjeta roja o amarilla.

Si te interesa jugar, entra en el siguiente link donde simplemente tenés que indicar nombre y dirección de correo y te enviaremos un mail con las instrucciones detalladas.

Ahora sí, a jugar....

Martín Elías Costa

2) **Instrucciones y reglas**

Estas instrucciones son formales y un poco largas pero el juego es muy sencillo y todo será evidente en cuanto empieces a jugar! Lee con atención los elementos los elementos en negrita que establecen algunos aspectos importantes del juego que no son auto-evidentes.

Objetivo del juego:

Existen dos tipos de jugadas: de una asociación o de tres asociaciones

En a jugada de una asociación vas a ver una palabra y tenés que escribir la primera palabra con la que la asocies, con las siguientes condiciones:

La palabra asociada no puede ser la misma que la que viste.
La palabra asociada ha de ser un sustantivo.
La palabra asociada no puede ser un nombre propio.

Si la primera palabra en que pensaste no cumple estas condiciones, escribí la siguiente que te venga a la cabeza.

Es importante jugar el juego correctamente, es decir escribir realmente la primer palabra en la que pensaste (siempre y cuando cumpla con las reglas anteriores) y no buscar hasta que encuentres una palabra "que te guste".

Una vez que hayas escrito la palabra tenés que elegir un jugador de la lista a quien querés enviársela. En algunas ocasiones tendrás la posibilidad de enviar esta palabra a dos jugadores. Este jugador

luego asociara esta palabra con otra, que será enviada a otro jugador y así siguiendo.

En la jugada de 3-asociaciones la consigna es la misma con la única diferencia de que tenés que escribir las tres primeras palabras con las que asocies la palabra presentada. Será fácil saber si una jugada es de una o tres asociaciones por el numero de casilleros disponibles para escribir las palabras asociadas. Es muy importante que pienses tres palabras asociadas a la palabra que se te presento y no hacer una cadena sucesiva de asociaciones. Para esto te pedimos que pienses las tres palabras y solo luego las escribas. Luego una de estas tres palabras será enviada al jugador que seleccionaste. Tenés que completar las tres palabras para seguir el juego.

Funcionamiento:

Si decidís jugar, entrá al link que esta abajo y llegaras a tu pantalla de juego. Ahí podes ver si tenés palabras pendientes que responder. La primera vez que entres tendrás una generada al azar. En las siguientes entradas a tu pantalla de juego, tendrás mas o menos palabras pendientes según si hayas o no recibido palabras de otros jugadores. Podes responder (asociar y enviar) al número que quieras de palabras de tu lista de palabras pendientes.

Podes entrar a tu pantalla de juego a ver si recibiste palabras y a responderlas en cualquier momento. Una o dos veces por semana, se te enviara un mail para avisarte si tienes palabras pendientes.

Para que el juego funcione correctamente es importante que las cadenas de palabras no se queden demasiado tiempo estacionadas en un jugador. A fin de evitar esto, de nuestra mejor consideración, otorgaremos tarjeta amarilla a todo jugador que en una semana no responda por lo menos a una cuota mínima de las palabras recibidas. Repetida esta circunstancia, el jugador será honrado con una suculenta tarjeta roja y un merecido agradecimiento por haber participado en el juego hasta ese punto.

Disclaimer:

Esta es una versión piloto del juego. Todo comentario, sugerencia, critica y observaciones será considerada y agradecida. A tal fin basta enviar un mail a juegodeasociaciones@gmail.com