

Dimensión de Espacios Métricos usando Fractales

J. G. Lalinde-Pulido

Ricardo A. Baeza-Yates

Depto. de Informática y Sistemas
Universidad EAFIT
Carrera 49 No. 7 Sur-50 A.A.3300
Medellín - Colombia
Fax: (574) 266-4284
email: jlalinde@eafit.edu.co

Depto. de Cs. de la Computación
Universidad de Chile
Blanco Encalada 2120
Santiago 6511224, Chile
E-mail: rbaeza@dcc.uchile.cl

Resumen

En este artículo introducimos un nuevo método para estimar la dimensión intrínseca de un espacio métrico arbitrario y lo comparamos con una definición previa que usa el histograma de distancias. Se incluyen resultados experimentales usando diccionarios y la distancia de edición, en dos lenguajes: inglés y español.

Palabras claves: algoritmos, bases de datos.

1 Introducción

Las bases de datos relacionales sólo permiten tipos de datos básicos, ya sea numéricos o strings de largo fijo, los cuales pueden ser consultados en forma exacta. Sin embargo, el número de aplicaciones con objetos complejos donde no es posible encontrar calces exactos aumenta cada día, en particular en bases de datos multimedia [2]. Un ejemplo muy concreto es el caso de imágenes y en particular impresiones digitales, donde es imposible encontrar la misma imagen por diferencias de escala, orientación, luminosidad, envejecimiento, etc. En estos casos se necesita buscar por el objeto más parecido a uno dado y es un problema de investigación con muchos resultados recientes [6]. Para definir que significa *más parecido*, la solución más sencilla es definir una función de distancia entre objetos y por ende obtener un espacio métrico. Es decir, un espacio donde la distancia es cero sólo al mismo objeto; donde la distancia es una función simétrica y no negativa; y donde se cumple la desigualdad triangular (es decir, $d(x, z) \leq d(x, y) + d(y, z)$).

En todos las estructuras de datos propuestas para buscar en espacios métricos se ha observado que el desempeño se degrada fuertemente cuando la dimensión de los datos es mayor, hasta tal punto que una búsqueda secuencial es más eficiente. Los ejemplos más importantes son BK-trees [4] y variantes como los FQ-trees [1], GNAT [3], M-trees [7], metric trees [9], entre otros. Esta limitación se ha llamado la *maldición de la dimensionalidad* y es fácil de observar en espacios cartesianos donde se puede controlar la dimensión. Vencer esta maldición es uno de los principales problemas abiertos en esta área [5]. En otros espacios más generales, si es posible definir una distancia, igual es difícil calcular la *dimensión intrínseca* del espacio. Por ejemplo, si tenemos tres puntos en un espacio tridimensional que representan una recta, la dimensión verdadera de esos tres puntos es 1 y no 3. Esto sería equivalente a la noción de dimensión intrínseca.

En [6] se propone estimar la dimensión intrínseca usando el histograma de distancias entre todos los pares de elementos del espacio usando la formula $d = \mu^2/2/\sigma^2$ donde μ es la media y σ la desviación estándar del histograma. En espacios cartesianos, esta noción coincide en orden con la dimensión real para un conjunto aleatorio de puntos. En este trabajo proponemos una definición alternativa basada en fractales para medir la dimensión intrínseca, la cual tiene una explicación geométrica que no es capturada totalmente en un histograma. Una comparación de las dos propuestas revela que no son equivalentes.

En nuestros resultados experimentales usamos el espacio de strings y la distancia de edición. La distancia de edición entre dos palabras a y b se define como el número mínimo de letras que tenemos que cambiar, borrar o agregar para convertir a en b . Este espacio además es discreto, lo que dificulta el cálculo de una dimensión, que es un concepto asociado a espacios continuos, y por ende le agrega un nivel de dificultad al problema. Nuestro método está adaptado especialmente a espacios discretos, pues en espacios continuos es inmediato.

Se propone entonces estimar la dimensión intrínseca del espacio analizando el conjunto de puntos mediante las técnicas utilizadas en fractales para estimar la dimensión de Hausdorff de un conjunto de puntos. La más utilizada es la denominada como dimensión *box counting*. La dimensión de Hausdorff se utiliza para caracterizar a los objetos fractales, pues en esos la dimensión de Hausdorff es mayor que la dimensión topológica. Dado que los objetos fractales son continuos y el diccionario es un conjunto discreto y finito, la estimación de la dimensión intrínseca del diccionario exige un tratamiento especial.

El artículo está organizado como sigue. Primero presentamos los conceptos básicos de dimensión fractal y luego nuestra propuesta para estimar la dimensión. Para diferenciar ambas propuestas, llamaremos DH a la propuesta anteriormente y BC a la nuestra.

2 Dimensión Fractal

Se propone estimar la dimensión intrínseca del espacio analizando el conjunto de puntos mediante las técnicas utilizadas en fractales para estimar la dimensión de Hausdorff de un conjunto de puntos. La más utilizada es la denominada como dimensión *box counting*. La dimensión de Hausdorff se utiliza para caracterizar a los objetos fractales, pues en esos la dimensión de Hausdorff es mayor que la dimensión topológica. Dado que los objetos fractales son continuos y el diccionario es un conjunto discreto y finito, la estimación de la dimensión intrínseca del diccionario exige un tratamiento especial.

La dimensión de Hausdorff basada en fractales describe la relación que hay entre la unidad que se utiliza para medir el conjunto y el valor de la medida obtenida. El proceso de medir el conjunto se refiere a recubrirlo con abiertos (esferas en el espacio métrico) de diámetro máximo δ . La unidad de medida es δ y el número de abiertos N necesarios para cubrir el conjunto es la medida obtenida [8].

2.1 Dimensión de Hausdorff

Sea $F \subset \mathbf{R}^n$. Diremos que $\{U_i\}$ es un δ -recubrimiento de F si $\forall i, |U_i| \leq \delta$ y

$$F \subset \bigcup_i U_i$$

Sea

$$\mathcal{H}_\delta^s(F) = \inf \left\{ \sum_i |U_i|^s : \{U_i\} \text{ es un } \delta\text{-recubrimiento de } F \right\}$$

La medida de Hausdorff s -dimensional de F se define como

$$\mathcal{H}^s = \lim_{\delta \rightarrow 0} \mathcal{H}_\delta^s(F) \quad (1)$$

y cumple que

$$\mathcal{H}^s(\Lambda F) = \Lambda^s \mathcal{H}^s(F) \quad (2)$$

Para cualquier F y $\delta < 1$ se tiene que $\mathcal{H}_\delta^s(F)$ es no creciente con s , y por lo tanto $\mathcal{H}^s$ también lo es. Sea $t > s$ y $\{U_i\}$ un δ -recubrimiento de F , entonces

$$\sum_i |U_i| \leq \delta^{t-s} \sum_i |U_i|^s$$

y por lo tanto

$$\mathcal{H}_\delta^t(F) \leq \delta^{t-s} \mathcal{H}_\delta^s(F) \quad (3)$$

Calculando $\mathcal{H}^s(F)$ y $\mathcal{H}^t(F)$ de acuerdo con 1 se observa que si $\mathcal{H}^s(F) < \infty$, entonces $\mathcal{H}^t = 0$ para $t > s$. Al analizar $\mathcal{H}^s(F)$ en función de s se observa que hay un valor crítico s' en el cual $\mathcal{H}^s(F)$ cambia su valor de ∞ a 0. Este valor s' se denomina dimensión de Hausdorff.

Si se define la dimensión topológica de un objeto como la dimensión del espacio de soporte, la dimensión de Hausdorff coincide con la dimensión topológica para objetos no fractales y es mayor que la dimensión topológica para fractales.

2.2 Dimensión de *Box Counting*

La dimensión de *box counting* es la técnica más utilizada de estimar la dimensión fractal de un conjunto de datos. Sea $\{U_i\}$ un δ -recubrimiento de F tal que $\forall i : |U_i| = \delta$. En este caso se tiene que

$$\sum_i |U_i|^s = N(\delta) \delta^s$$

donde N es el número de abiertos de diametro δ utilizados para recubrir F . Sea s la dimensión de Hausdorff de F , entonces podemos suponer que $N(\delta) \delta^s = 1$ sin perdida de generalidad normalizando la medida del objeto. Despejando s se obtiene

$$s = - \frac{\ln N(\delta)}{\ln \delta} \quad (4)$$

$$\ln N(\delta) = -s \ln \left(\frac{1}{\delta} \right) \quad (5)$$

Partiendo de este resultado, la dimensión de *box counting* se obtiene calculando $\ln N(\delta)$ y $\ln \left(\frac{1}{\delta} \right)$ para valores de δ cada vez más pequeños y calculando s' mediante una regresión lineal. El valor s' obtenido en la regresión es un estimador de la dimensión de Hausdorff real.

3 Estimación de la Dimensión Intrínseca

Para estimar la dimensión intrínseca del diccionario se utiliza la dimensión de *box counting* con unas modificaciones, las cuales se deben a la naturaleza misma del espacio conformado por el diccionario y la distancia de edición. Las principales características son:

- A diferencia de los objetos fractales, el diccionario es discreto y finito.

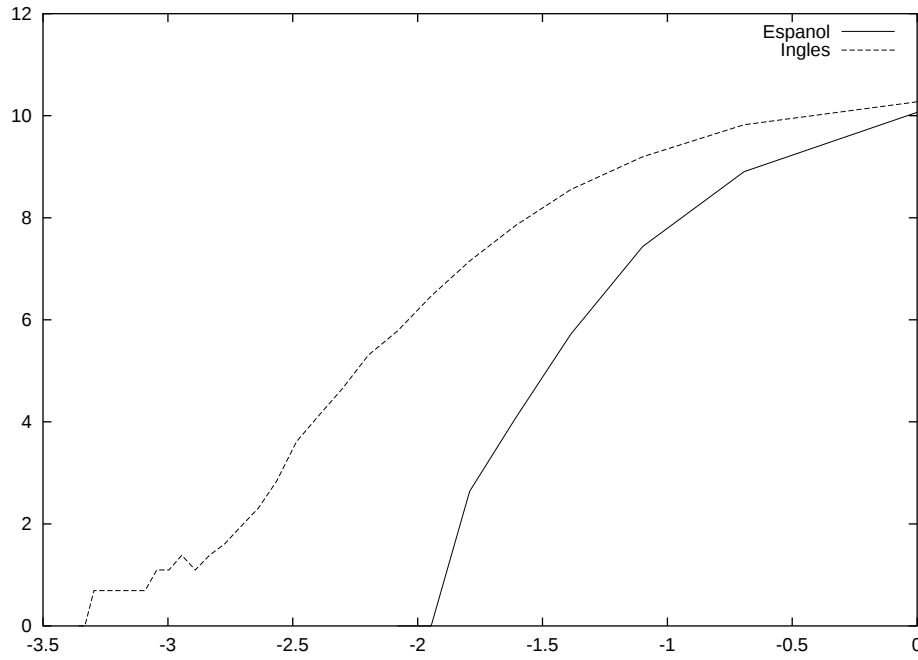


Figura 1: $\log N(\delta)$ contra $\ln\left(\frac{1}{\delta}\right)$

- La distancia es también discreta y acotada. Más específicamente, sólo puede tomar valores enteros y su valor máximo es la longitud de la palabra más larga del diccionario.

Estas características tienen una implicación que es muy importante: la dimensión de Hausdorff del diccionario es cero. Sin embargo, si se mira el diccionario como una muestra aleatoria de los puntos del espacio, el cual se supone además continuo, se puede estimar una dimensión intrínseca que sea útil para determinar la dificultad de las búsquedas.

Teniendo en cuenta estas características, si se grafica $\log N(\delta)$ contra $\ln\left(\frac{1}{\delta}\right)$ se debe obtener una gráfica de tipo sigmoideal, en la cual la pendiente es cero para $\delta < 1$ y para $\delta > D$ siendo D la distancia máxima entre palabras. Con el fin de estimar la dimensión intrínseca se realiza la regresión lineal teniendo en cuenta aquellos puntos que están en esa región intermedia.

En la figura 1 se aprecia la gráfica para un diccionario de inglés y otro de español. El eje x es $\ln\left(\frac{1}{\delta}\right)$, mientras que el eje y es $\log N(\delta)$. En la figura 2 se muestra la región usada para el diccionario de español y la regresión lineal usada para estimar la dimensión. Finalmente, en la figura 3 se muestra el histograma de las distancias para este último diccionario.

Ambas estimaciones fueron realizadas progresivamente desde un conjunto inicial pequeño del diccionario hasta el diccionario completo, para observar que dependencia hay, si es que la hay, respecto al número de elementos. Por lo que se observa en la tabla 1, para ambos diccionarios (español a la izquierda e inglés a la derecha), ambas definiciones no tienen las mismas variaciones, ni en dirección, ni en cantidad, a medida que el diccionario crece. Por otra parte, los valores para cada técnica no cambian mucho y el valor final en ambos casos indica que la dimensión del diccionario de español es mayor a la del inglés.

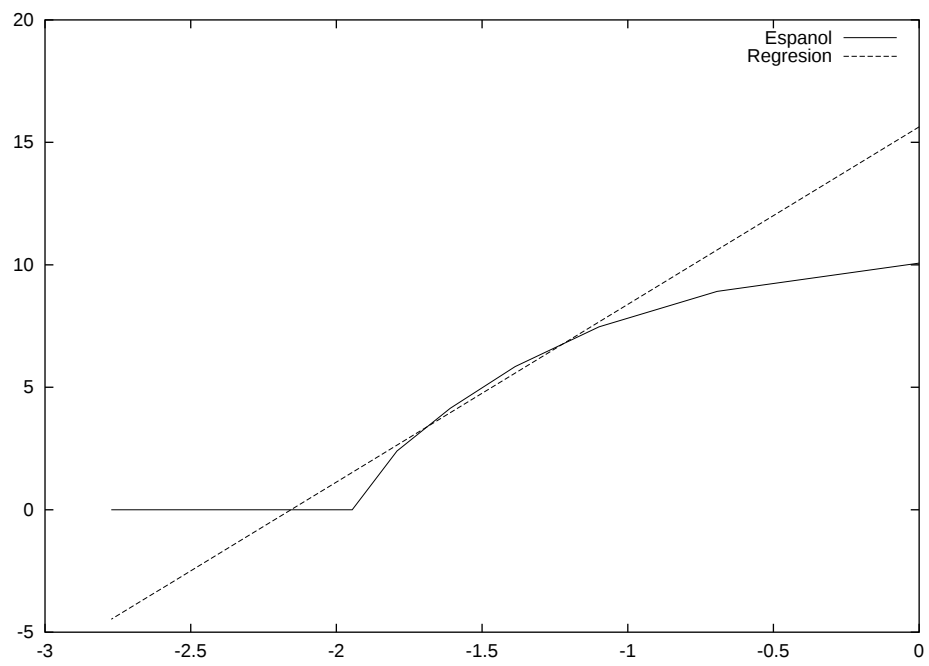


Figura 2: $\log N(\delta)$ contra $\ln(\frac{1}{\delta})$ y la regresión lineal usada.

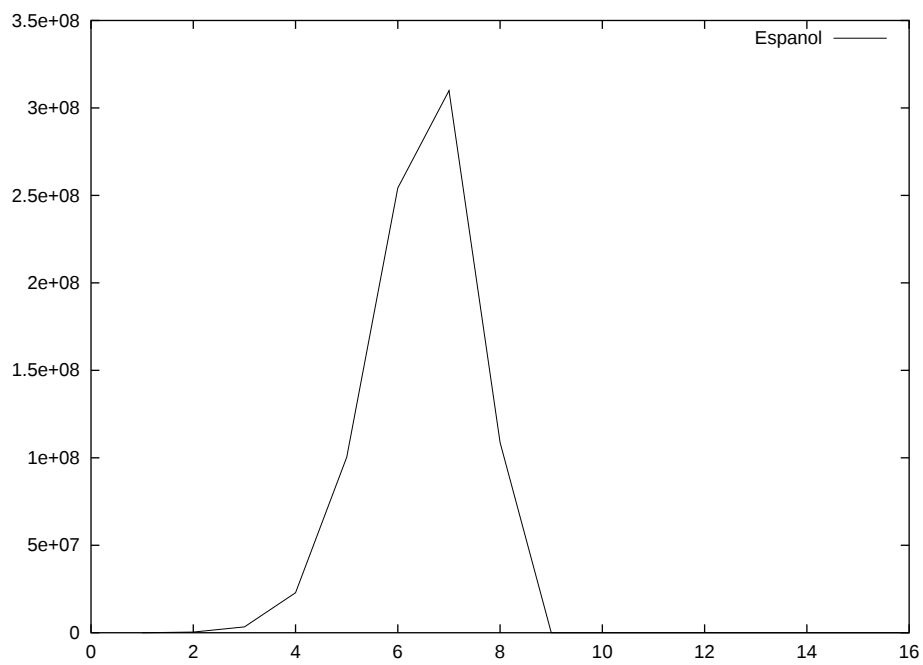


Figura 3: Histograma de distancia para el diccionario de Español.

Elementos	Método DH	Método BC	Elementos	Método DH	Método BC
4000	20.9542	5.4104	3281	8.4544	4.7122
4444	20.5703	5.6355	3645	8.0105	4.5301
5000	20.7733	7.2494	4101	8.1862	4.6229
5714	21.0196	6.6957	4687	8.0864	4.9065
6666	20.8312	5.7297	5468	8.2763	4.9549
8000	20.8111	5.7647	6562	8.4173	4.5666
10000	20.9194	6.2462	8203	8.1543	5.0917
13333	20.7381	6.3491	10937	8.2074	4.9151
20000	20.8974	6.3943	16406	8.1526	5.0095
40000	20.8397	6.5974			

Tabla 1: Resultados de ambos métodos para el español (izq.) y el inglés (der.).

4 Comentario Final

Nuestros resultados son aún preliminares y es necesario realizar más experimentos con otros conjuntos y otros espacios métricos. Por otra parte, estos resultados permiten calcular la dimension intrínseca de un espacio métrico, un problema que sigue buscando propuestas alternativas. Conocer la dimensión permite decidir mejor que estructura de datos es la más apropiada para una aplicación específica y por ende tiene valor práctico. En la versión final de este trabajo incluiremos más resultados para reforzar la propuesta aquí presentada.

Referencias

- [1] R. Baeza-Yates, W. Cunto, U. Manber, and S. Wu. Proximity matching using fixed-queries trees. In *Proc. 5th Combinatorial Pattern Matching (CPM'94)*, L NCS 807, pages 198–212, 1994.
- [2] R. Baeza-Yates and B. Ribeiro- Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.
- [3] S. Brin. Near neighbor search in large metric spaces. In *Proc. 21st Conference on Very Large Databases (VLDB'95)*, pages 574–584, 1995.
- [4] W. Burkhard and R. Keller. Some approaches to best-match file searching. *Comm. of the ACM*, 16(4):230–236, 1973.
- [5] E. Chávez, J. Marroquín, and G. Navarro. Overcoming the curse of dimensionality. In *European Workshop on Content-Based Multimedia Indexing (CBMI'99)*, pages 57–64, 1999. <ftp://garota.fismat.umich.mx/pub/users/elchavez/fqa.ps.gz>.
- [6] E. Chávez, G. Navarro, R. Baeza-Yates, and J. Marroquín. A survey of algorithms for searching in metric spaces. Technical Report TR/DCC-99-5, Dept. of Computer Science, Univ. of Chile, 1999. Submitted. <ftp://ftp.dcc.uchile.cl/pub/users/gnavarro/metricsurv.ps.gz>.

- [7] P. Ciaccia, M. Patella, and P. Zezula. M-tree: an efficient access method for similarity search in metric spaces. In *Proc. of the 23rd Conference on Very Large Databases (VLDB'97)*, pages 426–435, 1997.
- [8] Ricard V. Solé and Susanna C. Manrubia. *Orden y Caos en sistemas complejos*. Number 47 in POLITEXT Área de Física. Edicions UPC, 1996.
- [9] J. Uhlmann. Satisfying general proximity/similarity queries with metric trees. *Information Processing Letters*, 40:175–179, 1991.