Una Aplicación de Espacios Métricos en el Ámbito del Comercio Electrónico

Libertad Speranza, María de los Ángeles de la Torre, Norma Herrera

Departamento de Informática, Universidad Nacional de San Luis, Argentina,
libertadsperanza@gmail.com, angelesdlt@gmail.com,
nherrera@unsl.edu.ar

Abstract. El comercio electrónico (e-commerce) es un modelo de negocio que gestiona compra y venta de productos y/o servicios, y utiliza Internet como principal medio de intercambio. En este ámbito, los sistemas de recomendación son herramientas utilizadas para generar sugerencias sobre un determinado producto, En este trabajo abordamos el estudio de espacios métricos para la búsqueda de objetos similares en el entorno de un sistema de recomendación. Utilizamos para ello técnicas de indexación basadas en pivotes para resolver de manera eficiente la búsqueda de productos similares. Presentamos un estudio empírico de la eficiencia de estos algoritmos en este caso real de estudio.
Palabras claves: Espacios Métricos, Comercio Electrónico, Pivotes.

1 Introducción

Las operaciones de búsquedas en una base de datos requieren de algún soporte y organización especial a nivel físico. En el caso de las bases de datos clásicas, la organización de la información se basa en el concepto de búsqueda exacta sobre datos estructurados. Esto significa que la información se organiza en registros con campos completamente comparables (datos estructurados). Una búsqueda en la base retorna todos aquellos registros cuyos campos coinciden con los aportados en la consulta (búsqueda exacta).

Con el crecimiento de la web el volumen de datos e información que se maneja ha crecido exponencialmente, con lo cual las búsquedas se enfrentan a nuevos desafíos de rendimiento y velocidad de respuesta, dejando en segundo plano la congruencia de datos que ofrece el modelo relacional.

El comercio electrónico (e-commerce) es un modelo de negocio que gestiona compra y venta de productos y/o servicios, y utiliza Internet como principal medio de intercambio. Dentro de los desafíos que implica construir un sitio de e-commerce exitoso, sin duda uno de los más importantes es contar con un catálogo de productos bien organizado, donde el cliente no solo encuentre rápidamente lo que busca, sino que los resultados ofrecidos sean relevantes a lo que necesita.

En el campo del e-commerce, a menudo se necesita seleccionar una de varias alternativas sin tener un conocimiento exacto de cada una de ellas. La decisión final, en estos casos, puede depender de las recomendaciones realizadas por el propio sistema. La recomendación de productos a clientes es un caso de recuperación de información que no se puede abordar con la visión clásica de bases de datos relacionales.

En este trabajo abordamos el problema de buscar productos similares a un producto dado, obteniendo de esa manera recomendaciones para los usuarios de un sitio de e-commerce. Para ello aplicamos la teoría de espacios métricos [5], donde nuestro universo de datos será un conjunto de productos disponibles en la plataforma Mercado Libre, y la función que provee la medida de distancia, será la distancia de edición (o Levenshtein) aplicada al título principal de los productos. Este trabajo es la primera etapa de un proyecto mayor cuyo objetivo es construir un sistema de recomendación basado en el modelo de espacios métricos.

Este artículo está organizado de la siguiente manera: en la sección 2 y 3 describimos el marco de desarrollo de este trabajo dando un reseña de espacios métricos y de comercio electrónico. En la sección 4 presentamos nuestra propuesta, realizando la descripción y evaluación experimental de la misma en las secciones 5 y 6. Finalizamos en la sección 7 dando las conclusiones y el trabajo futuro.

2 El Modelo de Espacios Métricos

Nos interesa estudiar el problema de buscar con tolerancia en conjuntos de objetos. Dicho de otra manera, las búsquedas en donde se puedan recuperar objetos *similares* a uno dado. Para ello, abstraemos el problema a un universo de objetos $\mathcal{X}$ y una función de distancia d que modela la similitud entre los objetos del universo. Esta función d cumple con las propiedades características de una función de distancia: *positividad* ($d(x, y) \geq 0$), *simetría* ($d(x, y) = d(y, x)$) y *desigualdad triangular* ($d(x, y) \leq d(x, z) + d(z, y)$). El par $(\mathcal{X}, d)$ se denomina *espacio métrico* . La base de datos será cualquier subconjunto finito $\mathcal{U} \subseteq \mathcal{X}$.

En este modelo de bases de datos, una de las consultas típicas que implica recuperar objetos similares es la *búsqueda por rango*, que denotaremos con $(q, r)_d$. Dado un elemento $q \in \mathcal{X}$, al que llamaremos *query* y un radio de tolerancia r, una búsqueda por rango consiste en recuperar los objetos de la base de datos cuya distancia a q no sea mayor que r, es decir, $(q, r)_d = \{u \in \mathcal{U} : d(q, u) \leq r\}$.

El tiempo total de resolución de una búsqueda contiene tres términos, a saber: *T= #evaluaciones de* d $\times$ *complejidad(d) + tiempo extra de CPU + tiempo de I/O.* En muchas aplicaciones la evaluación de la función d es tan costosa que las demás componentes de la fórmula anterior pueden ser despreciadas. Éste es el modelo usado en este trabajo; por consiguiente, nuestra medida de complejidad será la cantidad de evaluaciones de la función de distancia d.

Claramente una búsqueda por similitud puede resolverse de forma ineficiente en tiempo $O(n)$ examinando exhaustivamente la base de datos $\mathcal{U}$. Para evitar esto, se preprocesa $\mathcal{U}$ y se construye un índice, diseñado para ahorrar cómputos en el momento de resolver una búsqueda. Básicamente, los algoritmos de indexación pueden clasificarse en dos grupos [5, 3, 7, 9, 8]: *basados en pivotes* y *basados en particiones compactas*.

En este trabajo estamos interesados en los algoritmos basados en pivotes. Este grupo de algoritmos construyen el índice basándose en la distancia de los objetos de la base de datos a un conjunto de elementos preseleccionados que llamaremos *pivotes*.Para ello, se seleccionan k pivotes $\{p_1, p_2, \ldots, p_k\}$, y se le asigna a cada elemento a de la base de datos, el vector o firma $\Phi(a) = (d(a, p_1), d(a, p_2), \ldots, d(a, p_k))$. Durante la búsqueda

se usa la desigualdad triangular junto con la firma de cada elemento para filtrar objetos de la base de datos sin medir su distancia a la query q. Dada una búsqueda $(q, r)_d$, se computa la firma de la query q, $\Phi(q) = (d(q, p_1), d(q, p_2), \ldots, d(q, p_k))$, y luego se descartan todos aquellos elementos a, tales que para algún pivote p_i se cumple que $\mid d(q, p_i) - d(a, p_i) \mid > r$.

Uno de los principales obstáculos en el diseño de buenas técnicas de indexación es lo que se conoce con el nombre de *maldición de la dimensionalidad*. El concepto de dimensionalidad está relacionado a la dificultad o facilidad de buscar en un determinado espacio métrico. La dimensión intrínseca de un espacio métrico se define en [5] como $\rho = \frac{\mu^2}{2\sigma^2}$, siendo μ y σ^2 la media y la varianza respectivamente de su histograma de distancias. Es decir que, a medida que la dimensionalidad intrínseca crece, la media crece y su varianza se reduce. Esto significa que el histograma de distancia se concentra más alrededor de su media, lo que influye negativamente en el desempeño de los índices.

3 Comercio Electrónico: Sistemas de Recomendación

La forma más conocida y popular de almacenar información se basa en las relaciones entre los distintos atributos (bases de datos relacionales). En éste tipo de universo, las consultas vienen dadas por una especificación de criterios que deben cumplirse para que un objeto forme parte del resultado; pero ¿qué pasa cuando en la consulta no podemos definir ese conjunto de criterios en forma determinística?

En el campo del e-commerce, a menudo se necesita seleccionar una de varias alternativas sin tener un conocimiento exacto de cada una de ellas. La decisión final, en estos casos, puede depender de las recomendaciones realizadas por el propio sistema. Los sistemas de recomendaciones [1, 6] son herramientas que generan sugerencias sobre un determinado objeto de estudio, a partir de las preferencias y opiniones dadas por los usuarios. El uso de estos sistemas va en aumento debido a que son muy útiles para evaluar y filtrar la gran cantidad de información disponible en la Web. El objetivo final es asistir a los usuarios en sus procesos de búsqueda y recuperación de información.

Un sistema de recomendación es un software que filtra información de interés para el usuario con el fin de proponerle aquel producto más adecuado a sus necesidades. Estos sistemas evalúan cuál es el grado de interés de un usuario por ciertos productos y buscan productos similares y con una alta probabilidad de atraer su atención.

Las recomendaciones pretenden ser una forma de ayudar al usuario a encontrar productos de su agrado realizando una preselección basándose, por ejemplo, en su historial de búsquedas. Esto proporciona un alivio a los consumidores, ya que les evita tener que recorrer una lista interminable de ofertas poco relevantes. Por otro lado, se espera que esto s beneficios para los usuarios se terminen reflejando en aumentos de tráfico y ventas, ya que en el e-commerce, las buenas recomendaciones siempre conducen a incrementos en las compras y, por ende, en los márgenes de ganancia. Es por ello que la efectividad del sistema de recomendación elegido es crucial.

El funcionamiento de un sistema de recomendación está basado en cierta información que es procesada por los algoritmos de filtrado. Dependiendo de la naturaleza de dicha información, podemos clasificar estos algoritmos en: basados en contenido, colaborativos, sensibles al contexto, entre otros.

En este trabajo nos hemos centrado en los algoritmos basados en contenido. Estos algoritmos filtran **objetos o contenidos similares** a los que el usuario ya ha buscado, comprado o calificado positivamente. Por ejemplo, en las plataformas de reproducción de música online, el software evalúa las piezas musicales analizando su estructura interna para encontrar piezas similares que podrían tener una línea de bajo parecida.

4 Espacios Métricos para un Sistema de Recomedación

En la plataforma de e-commerce Mercado Libre los usuarios publican productos dentro de un árbol de categorías; dichos productos constan de un título principal, un título secundario y una descripción.

El problema que abordamos en este trabajo es la búsqueda de los productos similares a un producto específico, obteniendo de esa manera recomendaciones para los usuarios. En este caso el universo de datos del espacio métrico está formado el conjunto de productos disponibles y la función de distancia será la distancia de edición (o Levenshtein) aplicada al título principal de los productos. Esta función de distancia es discreta y calcula la mínima cantidad de caracteres que hay que agregar, intercambiar y/o eliminar a una palabra para obtener otra. Este modelo es el más comúnmente usado en recuperación de texto.

El objetivo de este trabajo es, teniendo en mente un sistema de recomendación, encontrar una forma eficiente de resolver la búsqueda por similitud evaluando técnicas conocidas en este caso real de aplicación. Para lograr nuestro objetivo, primero organizamos los productos por categoría con el enfoque clásico de búsqueda exacta y luego dentro de cada categoría aplicamos el enfoque de espacios métricos. En consecuencia, la búsqueda de un producto implica primero filtrar por categoría realizando una búsqueda exacta y luego realizar la búsqueda por similitud sobre los productos que pertenecen a la categoría elegida.

Para el filtrado por categoría armamos un hashing lineal en memoria principal. Para la búsqueda por similitud utilizamos el enfoque basado en pivotes: elegimos k pivotes y para cada producto x almacenamos su firma $\Phi(x) = (d(x, p_i), \ldots, d(x, p_k)$ donde p_i es el i-ésimo pivote y d es la función de distancia de edición. Un punto crucial es la técnica de selección de pivotes usados para construir la firma de los productos. La forma en que los pivotes son seleccionados afectan el desempeño de un índice cuando se realiza una búsqueda [2, 5, 4]. En nuestro caso elegimos la técnica de selección incremental de pivotes [2] que ha demostrado ser eficiente en los espacios métricos normalmente usados por la comunidad científica que trabaja en búsqueda por similitud. Esta técnica la comparamos con la selección aleatoria de pivotes, para poder tener una base de comparación sobre su desempeño en este caso de estudio.

5 Evaluación Experimental

La evaluación experimental tuvo como objetivo analizar la eficiencia de los algoritmos basados en pivotes en este caso real de aplicación. Cabe señalar que hubo un exhaustivo trabajo previo de seteo de parámetros, dado que los valores que conmunmente figuran

Grupo	Cant. de Categorías	Rango del tamaño	Cant. de Pivotes
1	6	[1.034 - 15.964]	16, 32, 64, 128, 256
2	12	[19.032 - 46.530]	64, 128, 256, 512, 1024
3	8	[57.198 - 136,323]	256, 512, 1024, 2048
4	4	[16.7995- 213.578]	512, 1024, 2048, 4096

Table 1. Organización de categorías en grupos.

en los trabajos de espacios métricos no resultaban adecuados en este ámbito. Describimos a continuación cada uno de los aspectos considerados.

Base de Datos. Para la evaluación experimental se utilizaron productos ofrecidos en la plataforma de e-commerce Mercado Libre. Estos productos están clasificados dentro de un árbol de categorías, donde cada categoría reúne productos relacionados. Se obtuvieron alrededor de 2 millones de productos, distribuidos en 12.000 categorías hojas y 30 categorías iniciales. Para armar el hashing se utilizó la categoría inicial, dado que utilizar la categoría hoja restringía demasiado el universo de búsqueda produciendo que las respuestas no eran adecuadas a lo que el usuario pretendería como recomendación.

Elección de pivotes. El hecho de que la cantidad de elementos de las categorías estuviera distribuído en un rango muy amplio (de 1034 elementos para la categoría más pequeña a 213.578 elementos para la categoría más grande), nos obligó a segmentar dichas categorías en 4 grupos de acuerdo a su tamaño. En base a este mismo criterio, realizamos la elección de la cantidad de pivotes, teniendo en cuenta el porcentaje que los mismos representaban sobre el total de productos de cada categoría y la capacidad de memoria principal, dado que en esta etapa del trabajo el objetivo era diseñar el índice en memoria principal. La distribución resultante se presenta en la Tabla 1.

Otro punto a tener en cuenta era si se utilizaba el mismo grupo de pivotes para todas las categorías elegidos del universo total de búsqueda o cada categoría utilizaba su propio conjunto de pivotes elegidos dentro de esa categoría. Experimentalmente se estableció que la segunda opción era la mas adecuada y es la utilizada en los experimentos que se muestran en la próxima sección.

Selección del rango de búsqueda. La mayoría de los trabajos de investigación existentes sobre búsqueda por similitud en textos, utilizan como universo de datos diccionarios de palabras y como radio de búsquedas valores que van de 1 a 5. En nuestro trabajo, el universo de datos es bastante más singular, ya que se trata de títulos de productos reales, cuya redacción está a cargo del usuario que publica el producto para su venta y donde la única limitante es el tamaño de ese título (60 caracteres). Esta particularidad tiene como consecuencia un universo de datos variado y heterogéneo, donde cada elemento de dicho universo es una combinación de palabras, abreviaciones, números y caracteres especiales. Se realizaron una serie de experimentos previos que nos permitieron concluir que el radio de búsqueda mas adecuado para este caso es $r = 23$ (los resultados de estos experimentos no se muestran por razones de espacio).

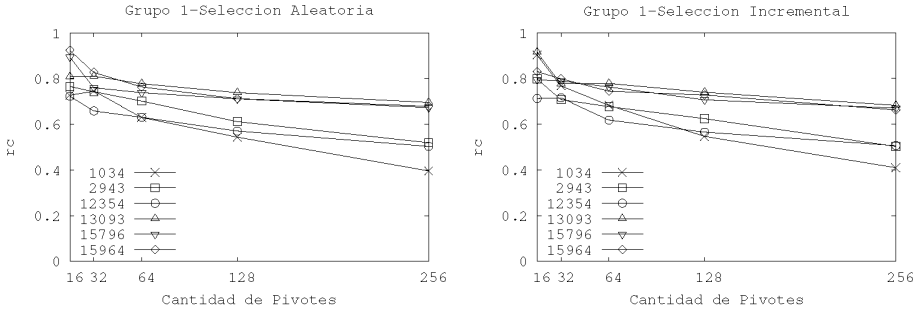

Fig. 1. Grupo 1: Efecto de la cantidad de pivotes sobre ***rc***.

6 Análisis de Resultados

6.1 Efecto de la Cantidad de Pivotes

Como ya mencionamos, el criterio de eficiencia del algoritmo de búsqueda, para cuando las estructuras se manejan en memoria principal, es la cantidad de evaluaciones de la función de distancia. En nuestro caso, como las categorías tienen diferentes cantidad de productos, para poder realizar una comparación global utilizamos el ***ratio de comparaciones(rc)*** definido como la cantidad de evaluaciones de la función de distancia dividido por la cantidad de elementos de la categoría. Es decir, ***rc*** nos da el porcentaje de elementos de la categoría con la que se comparó un elemento de búsqueda. De esta forma podemos obtener valores comparables independientemente de la cantidad de elementos que tenga cada categoría. El primer paso fue analizar el efecto de la cantidad de pivotes sobre ***rc*** para cada categoría considerada.

Para los expermientos, se seleccionaron al azar el 10% de los elementos de cada una de las categorías para realizar las búsquedas. Los resultados que se muestra son el promedio de las búsquedas realizadas. En las gráficas, cada categoría analizada queda identificada por la cantidad de elementos que contiene. Recordemos que las categorías están organizadas en grupos y para cada grupo hemos usado diferentes cantidades de pivotes. Los resultados que aquí se muestran reflejan esta organización. En el caso del grupo 2, por cuestiones de claridad, solo se muestran los resultados para 6 de las categorías de dicho grupo; las restantes categorías de ese grupo tienen un comportamiento similar a lo aquí mostrado.

En la Figura 1 observamos los resultados obtenidos con las categorías del Grupo 1. Sobre el eje x están representadas la cantidad de pivotes utilizados y sobre el eje y el ***rc***. Como era de esperar, tanto para la selección de pivotes aleatoria (izquierda) como para selección de pivotes incremental (derecha), el ***rc*** disminuye a medida que aumentamos el número de pivotes. Puede observarse que el mejor valor logrado para rc es cercano a un 70% para las categorías mas grandes y 40% para las categorías más chicas.

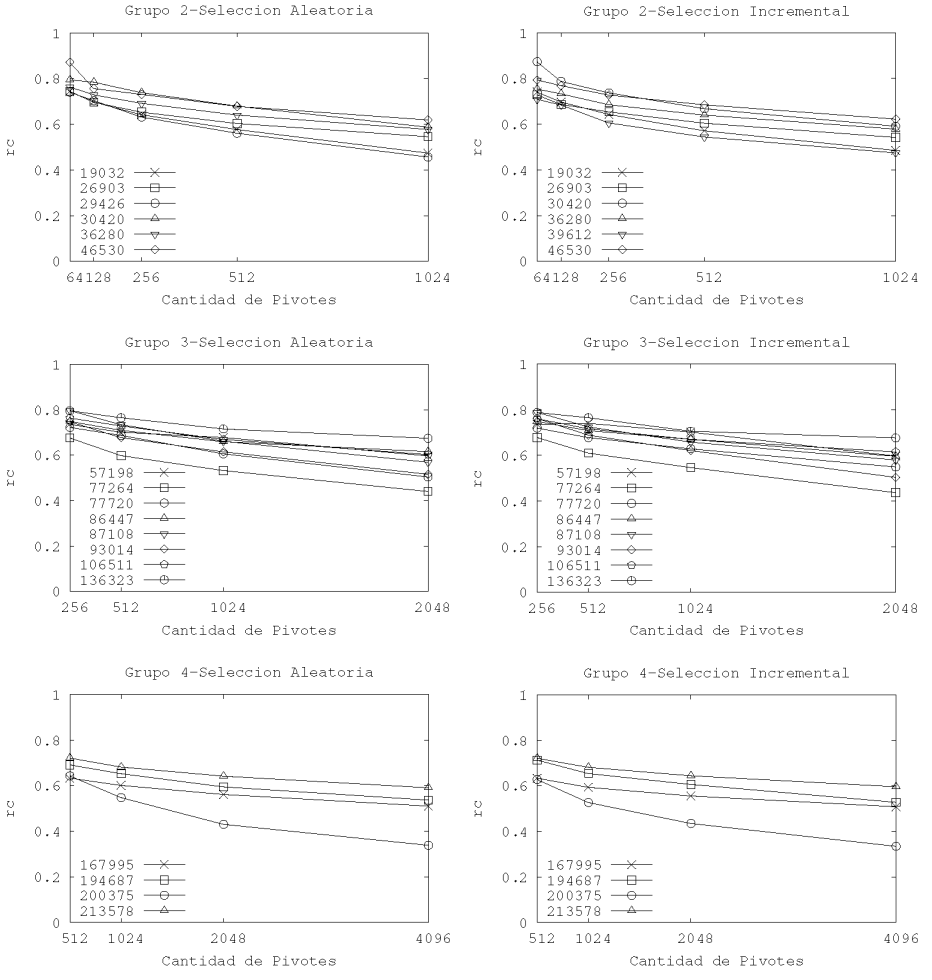

Fig. 2. Grupos 2, 3, y 4: efecto de la cantidad de pivotes sobre ***rc***.

Este mismo comportamiento se repite con los otros 3 grupos (ver Figura 2). En todos lo grupos de categorías el rc está entre un 40% y un 70% aproximadamente. La diferencia entre los distintos grupos radica en la cantidad de pivotes necesitados para lograr estos resultados: para las categorías más grandes (grupo 4) se necesitan 4096 y para las más chicas con 256 pivotes. Esto nos indica de un claro efecto del tamaño de la base de datos sobre los resultados obtenidos.

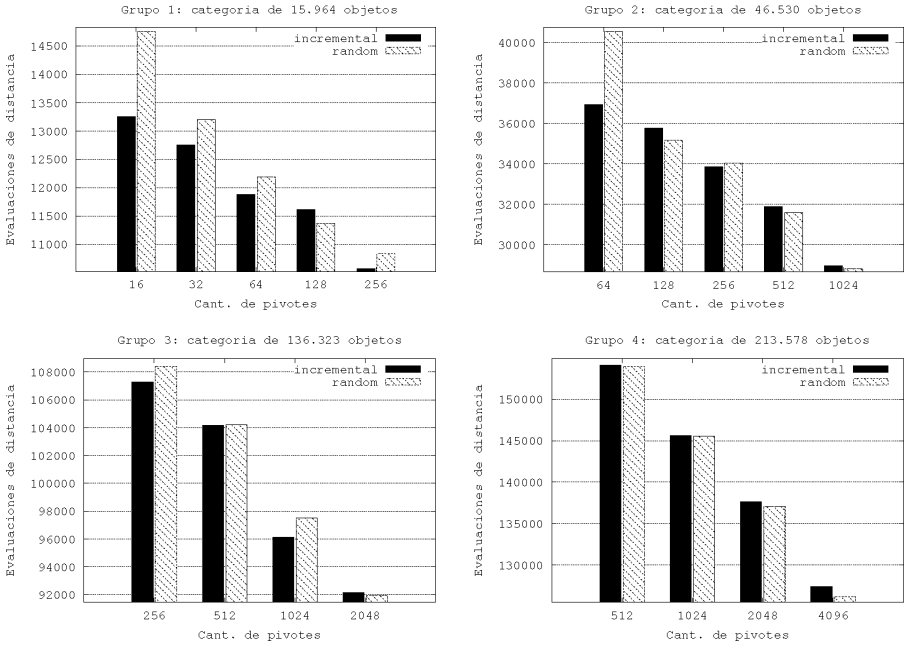

Fig. 3. Selección de pivotes: Incremental vs. aleatoria.

Podemos concluir que para todas las categorías consideradas a medida que aumentamos el número de pivotes mejoramos el ***rc***. Este resultado es el esperado cuando trabajamos con algoritmos de indexación basados en pivotes. La diferencia que encontramos respecto de lo que se encuentra en trabajos de investigación referidos a esta temática, es que la eficiencia lograda es menor para ambas técnicas de selección de pivotes.

6.2 Comparación de las Técnicas de Selección de Pivotes

Otro punto que se evaluó es si la técnica de selección incremental lograba mejorar a la selección random en la misma proporción que ocurre en los espacio métricos de prueba utilizados por la comunidad dedicada al estudio de espacios métricos. En la Figura 3 se muestran las gráficas para las categorías de mayor tamaño de cada grupo porque consideramos que son los más representativos; sobre el eje x está representada la cantidad de pivotes y sobre el eje y la cantidad de evaluaciones de la función de distancias hechas en la categoría considerada en cada caso.

Como puede observarse, en las categorías mas chicas (grupo 1 y 2), la técnica de selección de pivotes incremental tiene, en la mayoría de los casos, mejor desempeño que la selección aleatoria. Pero a medida que aumentamos el tamaño de las categorías

(grupo 3 y 4), esta diferencia disminuye llegando incluso a superar en eficiencia la selección random a la incremental. Por ejemplo, para la categoría más grande (213.578 elementos) con 4096 pivotes, la selección aleatoria hace aproximadamente un 50% menos de comparaciones que la selección incremental. Esto marca una diferencia de comportamiento de esta técnicas en este caso real de estudio respecto de lo que ocurre en los espacios métricos de pruebas normalmente utilizados en la temática.

6.3 Estudio de la Dimensionalidad

Los resultados obtenidos con la evaluación experimental fueron un indicativo de que podíamos estar ante un espacio de alta dimensionalidad. Dado que cada categoría es un universo completamente distinto, decidimos realizar el cálculo de los histogramas de distancia para todas las categorías consideradas. En la Figura 4 se muestran los histogramas de las categorías con mayor cantidad de elementos de cada grupo porque son las más representativas; sobre el eje x están representados los distintos valores de la función de distancia d y sobre el eje y la cantidad de elementos que están a esa distancia. Puede observar como se concentran los elementos alrededor de la media lo que indica que estamos ante un espacio de alta dimensionalidad, lo que influye negativamente en las búsquedas. En las secciones anteriores vimos que, independientemente de la técnica de selección de pivotes usada, el ***rc*** para la categoría más grande con 4096 pivotes ronda el 60%. Esto significa que ante una búsqueda, el algoritmo debe comparar con mas de la mitad de los elementos de la base de datos para obtener la respuesta a la consulta.

7 Conclusiones y Trabajo Futuro

En este trabajo hemos abordado el estudio modelo de espacios métricos aplicado a un caso real de estudio. Específicamente hemos usado el enfoque de indexación basado en pivotes para resolver la búsqueda de productos similares a uno dado, en un entorno de comercio electrónico. Nos propusimos estudiar el comportamiento de las búsquedas por rango utilizando las técnicas de selección de pivotes incremental y aleatoria. Podemos concluir que a medida que aumentamos la cantidad de pivotes mejora el comportamiento del índice, pero no en la misma proporción que ocurre en los espacios de prueba generalmente utilizados en esta tématica. También se pudo observar que la selección incremental no siempre mejora a la selección aleatoria. Como trabajo fututo nos proponemos estudiar el comportamiento de índices basados en particiones compactas y diseñar además un índice en memoria secundaria, para luego implementar un sistema de recomendación basado en espacios métricos.

References

1. Asim Ansari and Skander Essegaier. Internet recommendation systems. In *Journal of Marketing Research, ACM E-commerce 2000 Conference*, pages 17–20, 2000.
2. B. Bustos, G. Navarro, and E. Chávez. Pivot selection techniques for proximity searching in metric spaces. In *Proc. of the XXI Conference of the Chilean Computer Science Society (SCCC'01)*, pages 33–40. IEEE CS Press, 2001.

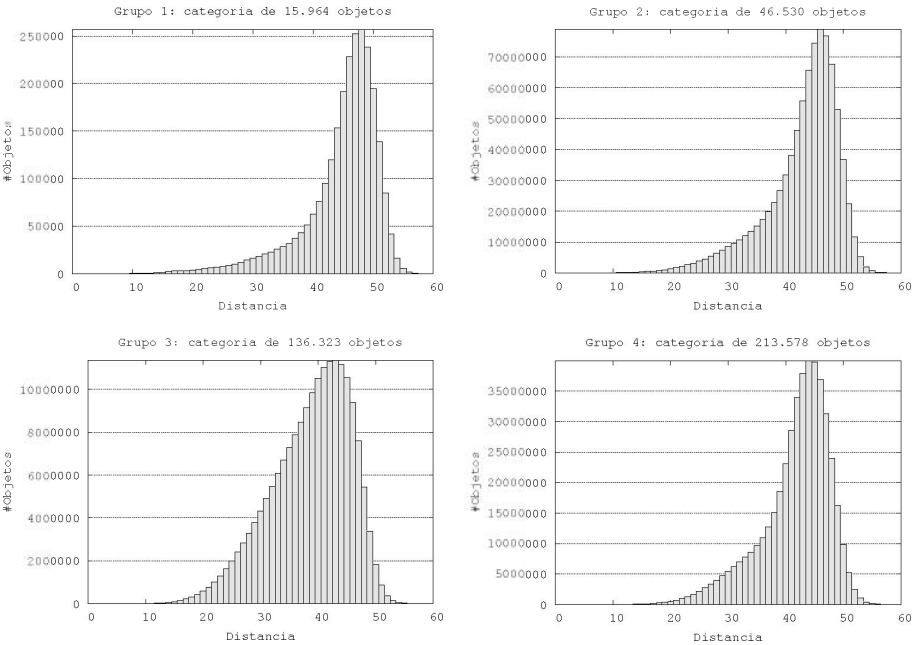

Fig. 4. Histogramas de distancias para las categorías mas grandes de cada grupos.

3. Benjamin Bustos and Nelson Morales. On the asymptotic behavior of nearest neighbor search using pivot-based indexes. In *Proc. 3rd International Conference on Similarity Search and Applications (SISAP'10)*, pages 33–39, New York, NY, USA, 2010. ACM.
4. Benjamin Bustos, Oscar Pedreira, and Nieves Brisaboa. A dynamic pivot selection technique for similarity search. In *Proc. 1st International Workshop on Similarity Search and Applications (SISAP'08)*, pages 105–112, 2008.
5. E. Chávez, G. Navarro, R. Baeza-Yates, and J.L. Marroquín. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, September 2001.
6. J. Franganillor. El funware en los sitios de recomendación social: características, claves e inconvenientes. In *: Anuario ThinkEPI: análisis de tendencias en información y documentación, vol. 12*, pages 242–250, 2018.
7. G. Navarro. Searching in metric spaces by spatial approximation. In *Proc. String Processing and Information Retrieval (SPIRE'99)*, pages 141–148. IEEE CS Press, 1999.
8. G. Navarro and N. Reyes. Dynamic spatial approximation trees for massive data. In *Proc. 2nd International Workshop on Similarity Search and Applications (SISAP)*, pages 81–88. IEEE CS Press, 2009.
9. G. Navarro and N. Reyes. New dynamic metric indices for secondary memory. *Information Systems*, 2016. To appear.