Concentración de la Distancia en Espacios Vectoriales de Alta Dimensión.
Concentration of Distance in High Dimensional Vector Spaces .

Director: *ALONSO, Juan Manuel*
Correo Electrónico: *jm31415ac@gmail.com*

Co-Director: *-*

Integrantes: *CALDERON, María Celeste.*

Palabras Clave: *Concentración de la Distancia, Dimensión Fractal Finita, Espacios de Palabras, Medidas de Similitud*

Resumen Técnico: *El objetivo general del proyecto es estudiar el fenómeno de concentración de la distancia en espacios euclidianos de alta dimensión. Estos espacios son una herramienta esencial para una gran cantidad de aplicaciones prácticas en informática, como por ejemplo bioinformática, análisis de imágenes, tecnología para acceder a información y redes de comunicación. Estos dominios de aplicación tienen en común que utilizan representaciones multidimensionales, y alguna noción de "similitud", en espacios vectoriales de alta dimensión. Generalmente, pero no siempre, "similitud" significa una distancia, en sentido matemático, en el espacio, siendo típica la utilización de una norma Lp, para algún 0 < p =< infinito. Tales modelos son, generalmente, fáciles de entender y sencillos desde un punto de vista calculatorio (aunque, por su tamaño, no siempre son fáciles de manejar). Sin embargo, la intuición que tenemos de la noción de distancia en el espacio tridimensional, no es suficiente para anticipar qué tipo de efectos calculatorios se pueden producir en estos modelos, que tiene una dimensión considerablemente mayor. Investigaciones recientes han demostrado que, potencialmente, hay un problema serio con la noción de distancia en espacios de alta dimensión. Se trata de ciertos conjuntos finitos que aparecen naturalmente en algunos de los dominios de aplicación mencionados, que tienen la propiedad de que su diámetro casi coincide con la distancia mínima entre dos puntos (distintos) cualesquiera del conjunto. En este caso, la noción de "punto más cercano", que es crucial en muchas aplicaciones prácticas, deja de tener sentido, ya que "cerca" y "lejos" es casi lo mismo. Este fenómeno, llamado de "concentración de la distancia", es de importancia capital en algoritmos que usan información de alta dimensión (en algunos dominios, como la investigación oncológica [2] este fenómeno ya ha sido identificado como un problema serio).Estudiaremos el fenómeno de concentración de la distancia desde un punto de vista geométrico, usando el concepto de dimensión fractal, y también el efecto que el uso de distintas distancias tiene sobre dicho fenómeno. Más aún, estudiaremos el fenómeno de concentración de la distancia para nociones de similitud que no sean necesariamente distancias en sentido matemático. Si bien la investigación es teórica, tenemos en mente un dominio concreto de aplicación: modelos de Espacios de Palabras. Se trata de modelos de representación de significado que se obtienen del lenguaje natural, de las palabras. Estos modelos son por su naturaleza de alta dimensión y, recientemente, han tenido mucho éxito en*

tecnología lingüística y búsqueda de información. Como ejemplo podemos citar a Gavagai AB (http://www.gavagai.se/) una empresa sueca creada en 2008 de la que el Dr. Alonso es consejero científico. Se propone la formación de recursos humanos con la incorporación de alumnos de la Licenciatura en Matemáticas.

Keywords: *Concentration of Distance, Fractal Dimension, Word Space, Similarity Measures*

Summary: *The overarching goal of this project is to investigate issues related to distances in high-dimensional vector spaces. More specifically, we will investigate concentration of distance from a geometrical point of view, with special emphasis on real-world data as encountered in natural language processing. We will also study how distance concentration affects the applicability of high-dimensional models and how the choice of different measures of similarity can redress (or exacerbate) the effects of distance concentration. We will develop criteria for identifying and diagnosing distance concentration in high-dimensional vector spaces. We expect the results of this project to be important and useful for any domain that uses similarities in high-dimensional vector spaces. We also expect the criteria and possible tools for diagnosing distance concentration to be useful both for research and practical applications.*