ACADEMIA DE CIENCIAS DE MORELOS, A.C.

¿Comentarios y sugerencias?, ¿Preguntas sobre temas científicos? CONTÁCTANOS:
edacmor@ibt.unam.mx

Física y Minería de textos o ¿qué investigan los cien

Jesús Antonio del Río Portilla
Centro de Investigación en Energía, UNAM-Campus Morelos (antonio@unam.mx)

Karla G. Cedano Villavicencio
Unidad de Difusión y Extensión, UNAM-Campus Morelos

Shirley Ainsworth
Instituto de Biotecnología, UNAM-Campus Morelos

No hay duda que **la** Internet ha cambiado muchos de nuestros hábitos de buscar información. Por ejemplo, en **la** actualidad basta tener alguna duda de cómo escribir una palabra para usar nuestro navegador y consultar el diccionario en línea de **la** Academia de **la** Lengua Española; hace algunos años lo normal era buscar en un diccionario esperando que fuera uno actualizado y que **la** palabra estuviera bien escrita o que no fuera muy rara para encontrarla. En cambio ahora con las búsquedas por aproximación lo primero ya no es imperioso y, con respecto a lo segundo, hasta tenemos **la** opción de buscar en **la** parte panhispánica de dudas del diccionario además de la historia lexicográfica..

En estos días es normal hacer una búsqueda y obtener miles de opciones y a veces quisiéramos tener **la** posibilidad de hacer un resumen o extraer los aspectos relevantes de toda esa información que nos brinda **la** Internet.

Un ejemplo de **la** avalancha de información que podemos recibir es **la** respuesta a **la** pregunta ¿Qué investigan los científicos en Morelos? Ya explicamos en estas páginas (http://www.acmor.org.mx/descargas/10_nov_08_cienciometria.pdf) que hemos encontrado más de diez mil artículos donde los científicos que laboran en Morelos han transmitido sus hallazgos. Claramente, para saber qué hacen, tendríamos que leer todos esos artículos. Aunado a esto, dado que un artículo de investigación explica nuevos hechos o descripciones como producto del trabajo de personas altamente especializadas en alguna rama del conocimiento, entonces **la** tarea de entender estas nuevas aportaciones, es mucho más complicada para las personas ajenas a esa área.

Precisamente, el entendimiento de estas aportaciones es lo que posibilita **la** aplicación de **la** ciencia y **la** tecnología en los ámbitos económicos, sociales y ambientales para conseguir el bienestar.

Ante esta situación es fácil caer en **la** tentación de aplicar encuestas o cuestionarios para determinar las líneas de investigación de las instituciones o de los científicos o tecnólogos. **La** investigación científica es cambiante, los científicos abordan diferentes problemas y están en una búsqueda constante de nuevas aportaciones. Las encuestas serían el retrato de un pasado y ni siquiera fiel, ya que frecuentemente el mismo investigador no ve algunas aplicaciones. Estas encuestas son equivalentes a solicitarle al científico un informe escrito de sus labores, es encargarle una tarea doble, el científico ya escribió y detalló sus hallazgos, están publicados, es más están en Internet, ¿por qué pedírselo nuevamente?

También en el artículo anterior (http://www.acmor.org.mx/descargas/10_nov_08_cienciometria.pdf) encontramos las 11 áreas de investigación más prolíficas de **la** mayoría de los artículos científicos, donde por lo menos un autor tiene dirección institucional en Morelos. Es más, conocemos los títulos de las revistas con más de cien artículos: "Salud Pública de México", "Ingeniería Hidráulica en México" (ambas publicadas en Morelos), además de "Physical Review E", "Faseb Journal" y "Journal of Bacteriology". Sin embargo, esta información no es lo suficientemente fina como para dársela a un empresario para que decida invertir en un nuevo negocio, o a un gobernante para que defina una política pública de largo alcance. Se requiere conocer con mayor detalle lo que hacen los científicos. Es decir necesitamos, al igual que en nuestra consulta en Internet, algo que nos ayude a encontrar lo esencial de toda esa información, para después, con mayor calma, escudriñarla y obtener lo que nos sea de mayor utilidad.

Con esta idea en mente hace algunos años empezamos a utilizar herramientas de **la** física estadística para encontrar aspectos relevantes del quehacer científico en el Estado de Morelos, entre otras aplicaciones. Sí lector, escribimos correctamente, física estadística. Esta rama de **la** física ha estudiado por muchos años sistemas compuestos por muchas partes, digamos muchas partículas, buscando obtener descripciones sencillas de un comportamiento global. Por ejemplo, para describir un gas, no es necesario conocer el comportamiento de todas sus moléculas. En lugar de usar billones de billones de posiciones y velocidades de todas las moléculas en movimiento del gas, **la** ley de los gases ideales solamente contempla la presión, volumen y temperatura y con estas tres variables puede describir el estado del gas. Eso lo aprendimos desde **la** secundaria en **la** Ley de los gases ideales.

En este trabajo ilustramos **la** habilidad de **la** física estadística de abordar problemas complejos de una forma simple. En particular puede ser aplicada al análisis de **la** información y extraer frases relevantes de textos de una forma simple, es decir hacer *minería de textos*.

Uno de los conceptos que más utilizamos en **la** física estadística es **la** entropía, que está relacionada con aspectos de orden y desorden en el mundo microscópico o con los estados accesibles del sistema en el mundo macroscópico. Elliot Montroll (http://en.wikipedia.org/wiki/Elliott_Waters_Montroll) mostró, entre otras muchas cosas, que **la** entropía está relacionada a **la** varianza de las cantidades que caracterizan sistemas complejos, en particular, **la** varianza nos indica **la** certeza que se puede tener sobre esas cantidades. **La** varianza es el cuadrado de **la** desviación típica o estándar (ver http://es.wikipedia.org/wiki/Desviacion_tipica#Interpretaci.C3.B3n_y_aplicaci.C3.B3n). En esta ocasión queremos analizar el orden o desorden de las palabras y para ello mediremos **la** varianza de la distancia entre **la** misma palabra. Esta distancia es el número de palabras entre palabras iguales. Si tuviéramos un texto ordenado tendríamos que las distancias entre una misma palabra serían iguales y la varianza sería cero.

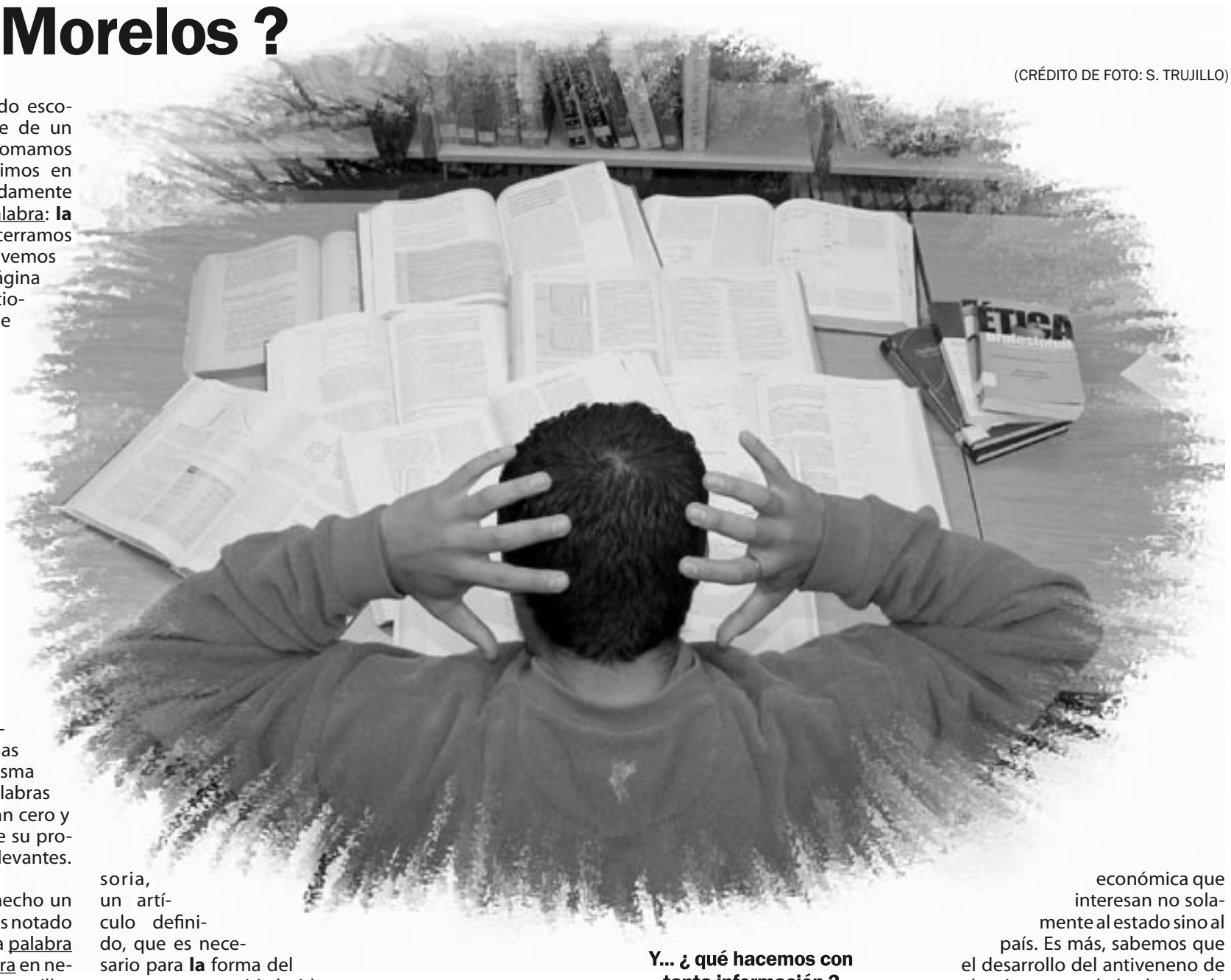
Detengámonos un momento en esta parte, ahora considere-

Alacranes, películas delgadas, corrosión...: algunos de los temas de investigación de los científicos de Morelos.

ACADEMIA DE CIENCIAS DE MORELOS, A.C.

¿Comentarios y sugerencias?, ¿Preguntas sobre temas científicos? CONTÁCTANOS: edacmor@ibt.unam.mx

tíficos en Morelos ?



(CRÉDITO DE FOTO: S. TRUJILLO)

mos un texto generado escogiendo aleatoriamente de un diccionario, es decir tomamos el diccionario, lo abrimos en alguna página y rápidamente seleccionamos una palabra: **la** ponemos en el texto; cerramos el diccionario y lo volvemos abrir en alguna otra página y nuevamente seleccionamos otra palabra que colocamos en el texto, así sucesivamente. Con este procedimiento tendríamos un texto completamente desordenado en sus palabras y las distancias entre las palabras tendrían varianzas del mismo tamaño que el promedio. De esta manera nuestro procedimiento para encontrar las palabras relevantes de los textos radica en calcular **la** varianza de las distancias entre la misma palabra y aquellas palabras cuyas varianzas no sean cero y que difieran mucho de su promedio son palabras relevantes.

En este texto hemos hecho un ejemplo, por eso habrás notado que hemos puesto una palabra subrayada y otra palabra en negrillas. **La** palabra en negrillas "**la**" tiene una varianza cercana a su promedio lo que quiere que decir es una palabra accesoria, un artículo definido, que es necesario para **la** forma del texto, no transmitir la idea y **la** vemos salpicada con desorden en este texto. En cambio **la** palabra "palabra" aparece colocada en algunos puntos claves, en lugares donde fue colocada intencionalmente es decir es una palabra relevante y su varianza es de 2.5 veces el valor de su distancia promedio indicando que no es una palabra colocada al azar ni en orden, sino que fue usada para explicar algo sustancial en este texto. Comentamos que otra palabra relevante en este texto es diccionario, le sugerimos al lector calcular la varianza normalizada de esta palabra.

(CRÉDITO: DISEÑO DE N. QUINTO)

Esta metodología nos permite encontrar las palabras o frases relevantes de textos y obtener información *'escondida'* en un mar de información sin necesidad de leerla y sin necesidad de ser especialista en el tema. En particular eso hicimos con los más de diez mil artículos de investigación de Morelos.

Durante varios años y en colaboración con Héctor Cortés y Jane Russell, colegas de **la** UNAM, hemos detectado algunas decenas de frases relevantes en los resúmenes de estos artículos. Algunas de estas frases relevantes son: *etli* (palabra ligada con un tipo de bacteria relacionada con el frijol), *CdS* y *CdTe thin films* (películas delgadas semiconductoras), *female commercial sex workers* (sexoservidoras), *Centruroides* (alacranes), *mean blood lead levels* (niveles medios de plomo en **la** sangre), *asthma* (asma), corrosión, toxinas. Claramente estas frases detallan con mayor precisión el espectro de las investigaciones en Morelos. De estas podemos inferir que los investigadores en Morelos abordan problemas sobre el campo como el frijol; películas delgadas semiconductoras para dispositivos fotovoltáicos; antídotos para piquetes de alacranes; problemas sociales íntimamente ligados con **la** salud; contaminación de plomo en los humanos; entre otros muchos tópicos de relevancia social y económica que interesan no solamente al estado sino al país. Es más, sabemos que el desarrollo del antiveneno de alacrán es uno de los logros de innovación basada en ciencia de Morelos. También se sabe que se tiene una compañía exitosa que protege las estructuras metálicas contra la corrosión. Estos son dos ejemplos de que las palabras detectadas pueden indicar con mayor precisión los nichos de innovación.

Y... ¿ qué hacemos con tanta información ?

La información anterior es un ejemplo de lo que se puede obtener con **la** minería de textos sin necesidad de leer los textos. Los resultados anteriores se obtuvieron en minutos al aplicar **la** física estadística y una herramienta computacional que implementa los algoritmos.

En resumen, **la** ciencia básica tiene repercusiones más allá de donde imaginamos y requiere que **la** sociedad, a través de su gobierno, apoye su desarrollo.

Agradecimiento: Queremos agradecer al Ing. Héctor Cortés por **la** implementación del algoritmo de minería en un código amigable.