

Proyectos de MinTIC

Los Nodos de Innovación son un espacio para pensar y después actuar. Los concibieron cuando las entidades del sector público salían al mercado en busca de soluciones de tecnologías y no las encontraban o estaban muy por encima de las posibilidades económicas.

Para encontrar una salida convocaron a la academia, al sector privado y a entidades y empresas del estado en sectores como justicia y salud, entre otros, y al Estado. Muchos tenían buenas ideas pero no las querían compartir. Ante ese hecho que no los dejaba avanzar, idearon las Convocatorias de los Nodos de Innovación. Invitaron a la academia y al sector privado a que se aliaran y presentaran proyectos de innovación cuyo objetivo fuera solucionar problemas en el Estado y la sociedad que podrían ser cofinanciados con los recursos públicos.

Dependiendo del caso, el MinTIC cofinancia hasta el 75 % del valor del proyecto, el cual debe generar un producto que se pueda explotar comercialmente.

Se han presentado más o menos 150 alianzas universidad-empresas, aunque solamente 32 han sido financiadas, porque muchos proyectos han tenido errores en su formulación.

Otro de los proyectos en marcha del MinTic es la conformación de un Tanque de Pensamiento que contribuya a la definición de la política pública.

De igual manera están siendo creados los Centros de Excelencia y Apropiación en los temas de *Big data analytics* e *Internet de las cosas*. Estos centros deberán ocuparse de formación, investigación, innovación y explotación comercial en estos temas, y para eso han sido conformados por universidades, empresas del sector TIC, y empre-

“Se debe avanzar para que el sector privado invierta en investigación y desarrollo”.

sas que esperan apropiarse de estas tecnologías en el futuro cercano. En el 2016 se dispone de 4000 millones de pesos para cada uno de estos centros y parte de la exigencia en la presentación de las propuestas fue mostrar un modelo de negocio que le permita al centro ser autosuficiente después de que se haya acabado el impulso inicial que les está dando el Ministerio. ■

Tecnología para combatir el tráfico humano y de armas

Sobre el desarrollo de un sistema de búsqueda por internet para localizar víctimas o traficantes de personas, así como compradores y vendedores ilegales de armas, habló Pedro Szekely, de la University of Southern California, en la conferencia “*Domain-Specific Knowledge Graphs*”.

En el tráfico de personas se mueven 32 billones de dólares al año en el mundo. La edad promedio de una “acompañante” es de 14 años. Sus explotadores ganan 150.000 dólares por cada una al año, lo que las obliga a atender diariamente numerosos clientes para completar esa suma.

Así lo explicó el ingeniero Pedro Szekely, quien expuso un sofisticado sistema de búsqueda entre los millones de avisos en internet para poder establecer relaciones encaminadas tanto a encontrar mujeres víctimas

de trata de personas como a sus explotadores. El sistema también ha sido utilizado para enfrentar otros delitos como el tráfico de armas y la venta de drogas ilegales.

El sistema desarrollado pretende ser mucho más preciso y afinado del que usan buscadores como Google, en cuyas bases de datos suelen estar solo los nombres de las personas y compañías más populares. “Google no es capaz de diferenciar entre muchas personas que tienen el mismo nombre. Por ejemplo, James Rodríguez hizo algo malo y un investigador policía

sabe que no es el mismo gran futbolista —explicó—. En internet hay otro James Rodríguez al que le gusta jugar fútbol pero también traficar en armas, pero el policía no puede encontrarlo porque Google decidió que el James Rodríguez que todo el mundo quiere encontrar es el que pensamos al oír el nombre”.

El sistema de búsqueda desarrollado por Pedro Szekely y su equipo también se emplea para encontrar a los *patents trolls*, compañías que compran patentes y después demandan o extorsionan a numerosas personas para que les paguen.

La dificultad para los investigadores es que cuando acuden a los buscadores populares para consultar los nombres de las *patent trolls*, en lugar de arrojar el listado de ellas, aparecen las declaraciones de la Corte Suprema de Estados Unidos referentes a ese problema.

Otro ejemplo de esas búsquedas fallidas o inexactas es el de la compra de armas en Estados Unidos. Adquirirlas es legal, pero hay ciertas restricciones. Sin embargo existen sitios donde las venden sin cumplir los requisitos. “Hay cientos de millones de avisos clasificados de armas en la web —explicó Pedro Szekely— y al buscar en Google ‘vendedores que no requieren *paper work*’, lo que sale son los artículos sobre las regulaciones”.

De ejemplos como esos surgió la idea de construir un sistema que permita a investigadores de la policía encontrar a estos vendedores o compradores.

Cómo funciona *Domain-specific Insight Graphs*

Teniendo en cuenta que es muy difícil combinar y relacionar tantos datos dispersos crearon DIG (*Domain-specific Insight Graphs*), un sistema de indexación, búsqueda y análisis para dominios específicos.

Cuando se está investigando a una persona desaparecida y se pone su nombre en internet aparece información dispersa en millones de páginas y hay que saber buscar. “No se pueden bajar todas las pá-

“ Cuando se está investigando a una persona desaparecida y en internet aparece información en millones de páginas y hay que saber buscar”.

ginas y parte del problema es identificar cuáles son las indicadas —explicó—. Una vez las bajamos, extraemos la información y así obtenemos datos estructurados; luego la vamos a mapear a una ontología. Después unimos todos esos *grafos* en un solo *grafo* coherente y sobre este hacemos las búsquedas”.

Para trabajar con 60 millones de páginas de avisos clasificados de mujeres acompañantes en internet, DIG cuenta con sistemas que hacen *crawling*: estos sistemas entran a la página, la bajan, se miran los enlaces y con heurísticas especializadas se decide si es interesante. De ser así, se emplean extractores que analizan el contenido y organizan su información. Algunos extractores se basan en la estructura HTML del sitio y trabajan a partir de ejemplos, utilizando algoritmos de *machine learning* para aplicar

los ejemplos y extraer información en muchas otras páginas.

Otros trabajan directamente sobre el texto en lugar de la estructura, pero tienen que lidiar con los intentos deliberados para ocultarle la información a buscadores tradicionales como Google. Por ejemplo, las páginas de acompañantes pueden hablar de “rosas” para referirse a “dólares” y pueden escribir el término como “ros4s”: es fácil de leer para una persona pero es difícil de entender para un buscador.

El siguiente paso es hacer un *mapping* de la información encontrada a una ontología diseñada para el dominio específico. “El problema es que DIG nos da datos en formatos diferentes y hay que ponerlos en la misma ontología para hacer las búsquedas”. Para ello usan Karma, un software (abierto) desarrollado hace cinco años al que se le dan la ontología y los datos que quieren estudiar. Este es flexible, escalable y facilita procesar miles de páginas por minuto.

Otra funcionalidad del sistema DIG se centra en las imágenes, pues muchos de los sitios de acompañantes o de vendedores de armas reutilizan una imagen cuando publican clasificados. Con el sistema intentan encontrar las fotos que no necesariamente deben ser idénticas, y para hallar dos similares entre 80 millones que han bajado de internet cuentan con colaboradores que usan tecnologías de *deep learning* para hacer búsquedas en segundos.

El sistema presentado ha tenido un gran impacto en los Estados Unidos, tanto en la comunidad de investigadores como entre las fuerzas del orden. Ya está siendo empleado en varios departamentos de policía, donde, basados en información encontrada, han abierto investigaciones y condenado traficantes. ■

<http://dig.isi.edu>



Pedro Szekely, de la University of Southern California.