

Nueva herramienta para procesar la ingente cantidad de textos de la web

Investigadores del centro CiTIUS de la Universidad de Santiago de Compostela han creado un software libre que acelera el procesamiento de textos y documentos publicados en la web. Su nombre es *Perldoop*, y permite el análisis de los datos de una forma más sencilla y eficiente.

CiTIUS - USC

25/11/2014 20:55 CEST



Los creadores de *Perldoop*. / Andrés Ruiz).

La ingente cantidad de información que se incorpora diariamente a Internet no para de aumentar. Se estima que en sólo 24 horas generamos aproximadamente 2,5 trillones de bytes (2,5 exabytes), o lo que es lo mismo: cerca de 27 GB por segundo, el equivalente a una temporada completa de *Juego de Tronos* en Alta Definición (HD). De hecho, el 90% de los datos disponibles actualmente en todo el mundo han sido creados apenas a lo largo de los dos últimos años.

De esta enorme cantidad de datos (agrupados bajo el anglicismo *Big Data*),

sólo el 5% se puede considerar información estructurada; el 95% restante (que está compuesto por textos, principalmente) no cuenta con ningún tipo de organización ni estructura, lo que representa un serio problema a la hora de acceder y gestionar toda la información disponible.

La herramienta adapta aplicaciones del ámbito del procesamiento de textos y documentos a modelos de computación

Ahora un equipo de investigadores del Centro Singular de Investigación en Tecnoloxías da Información ([CiTIUS](#)), formado por expertos en las áreas de Computación de Altas Prestaciones (HPC) y Procesamiento de Lenguaje Natural (PLN) de la Universidad Santiago de Compostela, ha desarrollado una herramienta que permite adaptar automáticamente aplicaciones utilizadas en el ámbito del procesamiento de textos y documentos a modelos de computación (en concreto a la computación paralela compatible con *clústeres multicore* o de multitud de nodos), lo que reducirá notablemente los tiempos de ejecución y permitirá trabajar con volúmenes de datos muy superiores a los que se manejan en la actualidad.

Los resultados obtenidos harán posible el análisis de los datos de una forma más sencilla y eficiente. Su propuesta se ha basado en el diseño de un nuevo sistema que permite transformar el software usado para el 'procesamiento del lenguaje natural' (PLN, habitualmente programado en el lenguaje informático *Perl*, y ejecutado de manera secuencial) en una solución compatible con las tecnologías *Big Data*.

Con sólo introducir unas etiquetas en la aplicación original, esta herramienta de traducción permite al programador convertir automáticamente todo su código *Perl* en código *Java* adaptado al denominado paradigma *MapReduce* (modelo de programación utilizado por Google para dar soporte a la computación paralela sobre grandes colecciones de datos), *habilitándolo* así para su ejecución en un *clúster*, es decir, permitiendo su ejecución simultánea en múltiples *cores* o nodos de computación.

De esta forma, se logra multiplicar la velocidad de cómputo por un factor

proporcional al número de procesadores disponibles (por ejemplo: si se dispone de 1.000 procesadores, el código resultante será, en el caso ideal, aproximadamente mil veces más eficiente que la solución secuencial).

***Perldoop*, una herramienta de código abierto**

Otra característica de esta investigación, que ha dado lugar a la herramienta de traducción [Perldoop](#), es que los resultados se han hecho públicos bajo una licencia de software libre, con el objetivo de que esté a disposición del mayor número posible de usuarios y profesionales. Como explica el profesor Juan Carlos Pichel, investigador principal y responsable del proyecto, la decisión se tomó porque "el desarrollo de soluciones *Big Data* para el PLN sólo está, en este momento, al alcance de las compañías más potentes". Con la solución propuesta, y unos conocimientos mínimos de programación, será posible convertir cualquier código programado en lenguaje *Perl* a una solución capaz de funcionar en un clúster de computación.

Entre las principales ventajas de esta nueva solución destaca especialmente su versatilidad, ya que se trata de una herramienta de propósito general; gracias a ello, podrán beneficiarse aplicaciones de ámbitos tan diversos como la traducción automática, el análisis de información en blogs, o incluso el procesamiento de datos genéticos.

Derechos: **Creative Commons**

TAGS

BIG DATA | WEB | INTERNET | COMPUTACIÓN | BIOINFORMÁTICA |

Creative Commons 4.0

Puedes copiar, difundir y transformar los contenidos de SINC. [Lee las condiciones de nuestra licencia](#)

