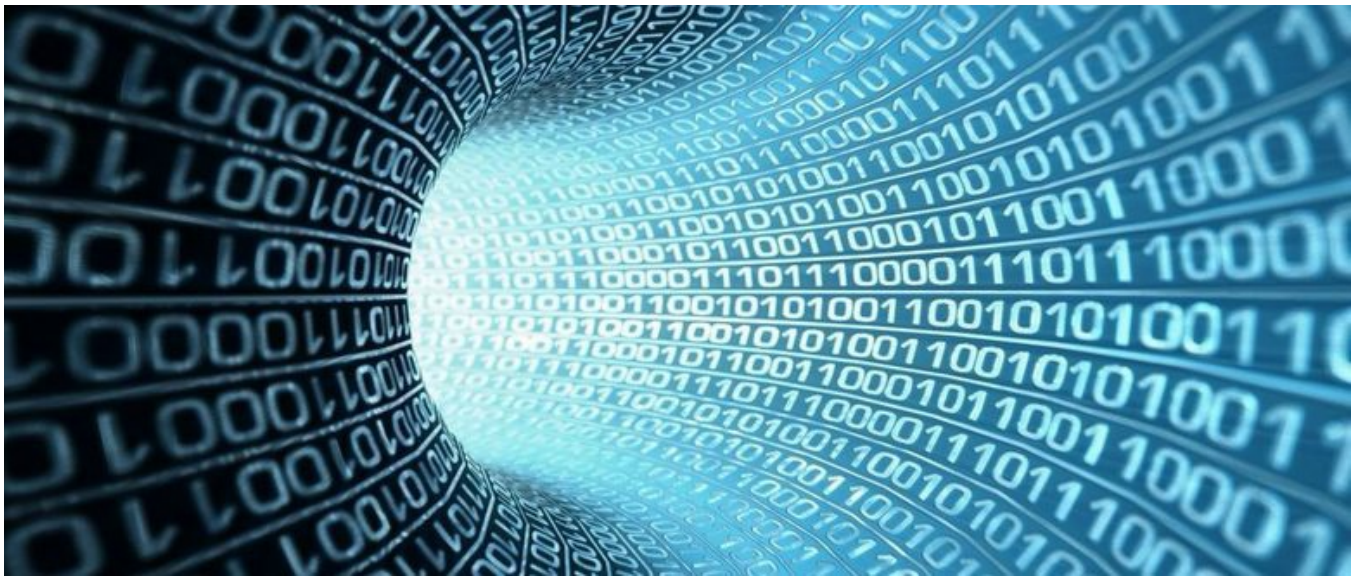
HPCNLP: Computación de Altas Prestacións para o Procesamento da Linguaxe Natural



Descrición

O Procesamento da Linguaxe Natural (PLN) é considerada como unha das metodoloxías máis apropriadas para poder estruturar e organizar a información textual accesíbel a través de Internet. O procesamento lingüístico de grandes cantidades de texto é unha tarefa complexa que require o uso de varias subtarefas organizadas en módulos interrelacionados. Un dos maiores problemas das técnicas de procesamento lingüístico é o seu alto custe computacional e a súas dificultades de escalabilidade, o que as fan inviábeis para a análise de grandes volumes (xigabytes e mesmo terabytes) de documentos. Deste xeito, o uso da Computación de Altas Prestacións (HPC) faise indispensabel se se quere reducir de forma notabel os tempos de cómputo, mellorar a escalabilidade do sistema, así coma no caso de querer abordar problemas dun tamaño aínda maior. Neste proxecto aplicaránse técnicas de paralelización/optimización e Big Data a diferentes prototipos que realizan diversas tarefas para o procesamento da linguaxe natural co obxectivo de integralos nunha suite de módulos PLN eficiente e escalábel. Os novos módulos PLN que se van desenvolver neste proxecto poderán utilizarse en aplicacións lingüísticas máis complexas e de alto nivel que verán así mellorar a súa eficiencia. Debemos destacar que as aplicacións de enxeñaría lingüística que poden beneficiarse destes módulos son: tradución automática, recuperación de información, busca de respostas, ou mesmo novos sistemas intelixentes de vixilancia tecnolóxica.

Obxectivos

As técnicas de PLN poden dividirse en dous grandes tipos de tarefas interrelacionadas: a análise do texto, por un lado, e a extracción de información, por outro. Os procesos de extracción utilizan, en xeral, texto analizado, e a ánalise textual mellora o rendemento cando se apoia en información previamente extraida do texto. No proxecto que propoñemos, aplicaremos estratexias de paralelización e optimización a tres tarefas específicas de PLN. Dúas destas tarefas se corresponden con métodos de análise lingüística: Recoñecemento de entidades con nome (NER - Named Entity Recognition) e Análise sintáctica de dependencias. A terceira tarefa que abordaremos forma parte das técnicas de extracción de información: Extracción de relacións semánticas entre entidades.

INVESTIGADORES

Investigador principal
Juan Carlos Pichel Campos

Investigadores do CiTIUS
Tomás Fernández Pena
Pablo Gamallo Otero
Marcos Garcia González

DETALLES

Data de execución:
08/08/2013 - 31/12/2014

Páxina web
http://proxectos.citius.usc.es/hpcpln/

Financiado por
Investigadores Emerxentes, Xunta de Galicia, Consellería de Educación e Ordenación Universitaria, EM2013/041

PUBLICACIÓNS

Yet another suite of multilingual NLP tools
Third Symposium on Languages, Applications and Technologies, 2015

Perldoop: Efficient Execution of Perl Scripts on Hadoop Clusters
IEEE International Conference on Big Data, 2014

Overview of TweetLID: Tweet Language Identification at SEPLN 2014
Twitter Language Identification Workshop at SEPLN 2014, 2014

Ver todas

PROGRAMAS CIENTÍFICOS

Computación avanzada