RECOPILACIÓN DE DATOS EN LÍNEA PARA EL REFINAMIENTO DE TRADUCCIONES AUTOMÁTICAS, UTILIZANDO LA WEB COMO BASE DE CONOCIMIENTO

FELIPE ANDRÉS VALDEBENITO SANDOVAL
INGENIERO CIVIL EN COMPUTACIÓN

RESUMEN

En esta memoria se propone el diseño de un Crawler capaz de recopilar de forma automática frases en español desde la Web, con el propósito de utilizarlas para el remamiento basado en frases de traducciones automáticas. El Crawler construido se basa en la arquitectura y comportamiento de Web Crawlers, los cuales recolectan información de páginas Web para máquinas de búsqueda. Para recorrer la Web en busca de frases se consideraron aspectos de la misma tales como su gran tamaño y rápido cambio de contenidos.

El diseño propuesto utiliza técnicas de computación paralela y distribuida para realizar la planificación URLs, recolección automática de frases en Español y almacenamiento de información. Se presenta un algoritmo para detectar el idioma del texto. Dicho algoritmo se basa en la cantidad de Stopwords que posee el texto para realizar la detección. Además, se realizo un breve estudio el cual permite determinar la cantidad de Stopwords que se utilizan en el idioma Español, mejorando la precisión en el proceso de detección de idioma.

El Crawler construido es capaz de descargar varias páginas Web al mismo tiempo, lo que puede provocar saturación en los servidores Web. Con la finalidad de evitar este problema se diseño un algoritmo de planificación de URLs por visitar. Finalmente, se evalúa el desempeño del Crawler construido. Además, se menciona el trabajo futuro que puede surgir de este trabajo.

ABSTRACT

This project shows the design to propose that a Crawler may automatically collect phrases in Spanish from the Web, in order to use the re_nement based on an automatic translation of the phrases. The Crawler is constructed based on the architecture and behavior of Web Crawlers, which collect information from Web pages for search engines.

To visit the Web and obtain phrases, there are many considered aspects, such as its large size and rapid change of content. The proposed design uses techniques of parallel computing, to plan and perform the automatic collection of phrases in Spanish, and store it to the hard drive. An algorithm was used to detect the language of the text. The algorithm is based on the amount of Stopwords to make the detection. A brief study was performed to allow us to determine the number of Stopwords used in the Spanish language. The built Crawler is able to download many Web pages at the same time, which can generate saturation in the Web server. To avoid this problem an algorithm was designed for the plani_cation of the URLs. Finally an evaluation was performed of the Crawler constructed. In addition, it is mentioned the work that may arise based on the work that was realized.