Universidad de Concepción - Chile

Facultad de Ingeniería

Departamento de Informática y Ciencias de la Computación

Detección y Búsqueda de Noticias Basado en Información Temática, Temporal y Espacial

Simón Cristóbal Smith Bize

Profesor Guía

María Andrea Rodríguez Tastets

Tesis presentada a la

ESCUELA DE GRADUADOS

DE LA UNIVERSIDAD DE CONCEPCIÓN

Para optar al Grado de

MAGÍSTER EN CIENCIAS DE LA COMPUTACIÓN.

Marzo, 2006

Resumen

El fuerte crecimiento de la World Wide Web ha hecho más difícil la tarea de reconocer y recuperar información de interés para un usuario. Los métodos clásicos de recuperación de información para analizar textos no o semi-estructurados, en base a las ocurrencias de las palabras claves dentro de un documento, tienen limitantes que motivan el estudio de nuevos métodos que exploren la semántica del contenido en documentos textuales. Esta Tesis explora la combinación de métodos tradicionales de recuperación de información con el análisis de contenido espacial y temporal de documentos para lograr la detección de tópico y recuperación de noticias en la Web chilena. La detección de un tópico se refiere a la tarea de construir clusters de noticias que discutan el mismo tema.

El trabajo propone analizar el documento de una noticia separando su texto en tres componentes: términos en el título, términos relevantes en el texto (p. ej. nombres propios) y los restantes términos en el documento. El componente temporal de una noticia es definido como su tiempo de publicación, donde se asume que una mayor cercanía temporal apoya la relación temática entre noticias. Similarmente, la referencia espacial de una noticia es asociada a la referencia geográfica, usualmente identificada en el encabezamiento de una noticia, y suponiendo que la cercanía geográfica apoya la relación temática de una noticia.

Esta Tesis pretende complementar trabajos previos que incorporan parcialmente el tiempo en la detección tópicos, utilizando algoritmos de clustering jerárquico de tipo single-link. El sistema es comparado con un sistema actual de detección de tópico (TDT) obteniendo mejores resultados de rendimiento. Adicionalmente, los resultados del algoritmo de detección de tópico son utilizados en un proceso de recuperación de información como método de indexación que es capaz de detectar noticias relevantes a una consulta con una menor dependencia del uso, en la especificación de la consulta, de términos relevantes en los documentos. Estos resultados complementan un proceso de navegación que permite recorrer noticias asociadas a un tópico de interés.