Instituto Politécnico Nacional

Escuela Superior de Computo

Recolector y Clasificador de Noticias

2018-B013

T E S I S

QUE PARA OBTENER EL GRADO DE:

LIC. EN INGENIERÍA EN SISTEMAS COMPUTACIONALES

PRESENTAN:

CARLOS ANDRES HERNANDEZ GOMEZ
LUIS DANIEL MEZA MARTINEZ

DIRECTORES:

M. EN C. JOEL OMAR JUÁRES GAMBINO
DRA. CONSULO VARINIA GARCIA MENDOZA

Índice general

1. Introducción **V**
1.1. Problemática VI
1.2. Solución Propuesta VI
1.3. Objetivo VI
1.4. Objetivos Específicos VI
1.5. Estructura del Documento VII

Capítulo 1

Introducción

La noticia es la información de un hecho de interés ocurrido durante un periodo de tiempo determinado. Constituye el elemento primordial de la información periodística y el género básico del periodismo [1]. Conocer los acontecimientos del mundo independientemente del tema, día, lugar en que se suscitan, tiene una gran importancia en la sociedad, estos se comparten por distintos medios de comunicación, tales como la televisión, redes sociales, diarios, blogs y la radio. Las noticias nos permiten conocer la situación económica del país, logros de la ciencia, desastres naturales, la situación en cuestión de inseguridad y otros acontecimientos. En el ámbito de las inversiones, las noticias crean expectativas y eso a su vez puede modificar los planes de inversión en cualquier sector, siendo así de suma importancia compartirlas de una forma eficaz [2].

Mucha de la información que se consulta hoy en día se hace a través de páginas web. Estas páginas son accesibles gracias a una herramienta llamada crawler que las reúne para indexarlas y hacer más sencillo que los motores de búsqueda puedan recuperarlas [3]. En la actualidad las páginas web van incrementando día con día, por lo cual se pueden consultar noticias de distintos sitios, alguno de estos son los periódicos electrónicos, los cuales dividen sus artículos en secciones para facilitar la búsqueda del usuario, sin embargo, el nombre de las secciones no coincide en todos los periódicos a pesar de que el tipo de contenido sea el mismo. Existen un sinfín de sitios independientes en la red, que proveen una gran variedad de artículos, dichos sitios no cuentan con una clasificación particular, por lo que resulta difícil para el usuario realizar una búsqueda específica dentro de dichos sitios. Dada la gran cantidad de sitios web que publican noticias, se han creado algunas aplicaciones similares a la propuesta en este trabajo que permiten la recolección de noticias de interés para el usuario como Flipboard [4], Huffpost [5] y Google News [6]. En la Tabla 1 se muestran dichas aplicaciones con sus características más relevantes.

1.1. Problemática

Hoy en día existen distintas maneras de informarse acerca de los acontecimientos más recientes, por ejemplo, en la televisión, blogs, redes sociales, foros, diarios, etc. Esto ha provocado que la información se encuentre más dispersa y se tenga que acceder a muchos recursos para recopilarla. Esta situación implica un gran esfuerzo y tiempo. Para ayudar en este problema existen herramientas que hacen la búsqueda de noticias de interés para el usuario en forma automática. Sin embargo, dichas herramientas dependen de que los sitios a consultar cuenten con una etiquetación correcta y homogénea de la información.

Según El Economista [8] el sitio web "Animal Político" ocupa el lugar número cuatro en el ranking de medios nativos digitales y clasifica sus noticias de una manera poco habitual para los lectores, como la sección "El sabueso", "El plumaje", "Hablemos de ... ", entre otras, lo que hace complicado obtener los artículos para los métodos tradicionales de recopilación que se basan sólo en las etiquetas que identifican cada sección y no el contenido de las noticias. Debido a lo anterior se propone crear un recolector de noticias el cual permita recopilar noticias de distintas fuentes de información, y mediante el análisis automático de su contendido determine si este guarda relación con las secciones de interés del usuario y el periodo establecido. Finalmente, las noticias que satisfagan ambos filtros serán las que se le mostrarán al usuario.

1.2. Solución Propuesta

Crear el TT1.

1.3. Objetivo

Crear un recolector de noticias, el cual permita recopilar información de diferentes fuentes como diarios, sitios de noticias, blogs, foros y mediante el análisis automático de su contenido muestre aquellas noticias que satisfagan los filtros de período y secciones establecidos por el usuario.

1.4. Objetivos Específicos

- Desarrollar un recolector de noticias, el cual permita recopilar información de diferentes fuentes como diarios, sitios de noticias, blogs y foros
- Analizar de forma automática el contenido de las noticias para satisfacer los filtros establecidos por el usuario
- Mostrar el enlace (URL) de las noticias que cumplieron con los filtros establecidos

- Afinar el clasificador de noticias realizado en el trabajo terminal 2017-A02 para utilizarlo en el contexto de esta propuesta (filtro de sección)

1.5. Estructura del Documento