Recolector y clasificador de noticias basado en su contenido.

Trabajo Terminal No. ———-———

Alumnos: Hernández Gómez Carlos Andrés, Meza Martínez Luis Daniel
Directores: Juárez Gambino Joel Omar, García Mendoza Consuelo Varinia
Turno para la presentación del TT: Matutino
e-mail: ldanielmezam@gmail.com

Resumen – En el presente trabajo terminal se propone desarrollar un recolector (crawler) de noticias, que permita recuperar artículos publicados en diferentes sitios de información. El recolector permitirá establecer dos filtros: el periodo en el cual se publicaron las noticias y la sección a las cuales pertenecen, por ejemplo, cultura y deporte. Los métodos tradicionales de recolección de información se basan en las etiquetas o marcadores que los sitios web añaden a sus páginas, en el caso de las noticias una de estas etiquetas puede ser la sección a la cual pertenecen estas. Sin embargo, existen fuentes de información que no dividen sus noticias en secciones o el nombre de las secciones no indica claramente el tipo de contenido, por ejemplo, la sección de deportes se pueden llamar adrenalina. Todo lo anterior complica la recuperación de la información. Por lo tanto, se propone que las noticias recuperadas sean clasificadas en secciones de forma automática de acuerdo con su contenido. Finalmente, todas aquellas noticias que cumplan con los filtros establecidos por el usuario serán mostradas para que puedan ser consultadas.

Palabras clave – Aprendizaje automático, clasificación de texto, web crawler.

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.

Características	Flipboard	Huffpost	Google News
Consulta diferentes fuentes	**Si**	**Si**	**Si**
Permite indicar secciones de interés	**Si**	**Si**	**Si**
Permite establecer periodo de interés	**No**	**No**	**No**
Clasifica los artículos automáticamente	**No**	**No**	**Si**

Tabla 1. Resumen de aplicaciones similares.

La diferencia más importante del trabajo propuesto con las aplicaciones similares, además del filtro del periodo de interés es el hecho de clasificar la información de forma automática para determinar a qué sección pertenece su contenido. Cabe mencionar

que existe un trabajo terminal realizado previamente titulado “Clasificación de noticias de diarios de circulación nacional mediante aprendizaje automático” con número 2017-A02 [7], en el cual se llevó acabo la clasificación de noticias en forma automática, y el trabajo propuesto tomará como base dicho clasificador para afinarlo y adaptarlo a la tarea del recolector de noticias.

2. 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.

Objetivos particulares

- 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).

3. Justificación

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” (www.animalpolitico.com) 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.

4. Productos o resultados esperados

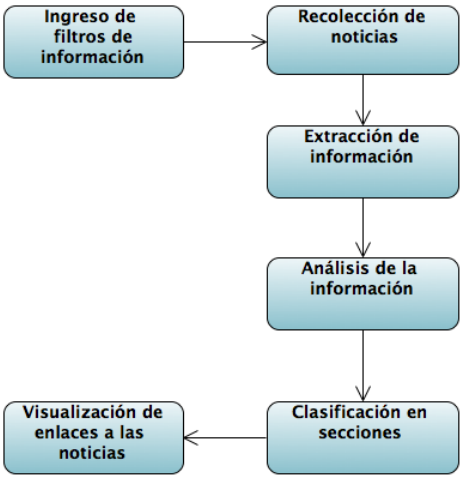

Figura 1. Arquitectura del sistema.

El sistema propuesto contará con los módulos mostrados en la Figura 1. A continuación se describen brevemente cada uno de ellos.

- Ingreso de filtros de información. Desde este módulo el usuario podrá establecer el periodo y las secciones de las noticias de su interés.
- Recolección de noticias. El recolector accederá a las fuentes de información registradas, recopilando la información publicada por estas.
- Extracción de información. En este módulo se procesará la información recolectada con el fin de extraer el contenido de las noticias.
- Análisis de la información. Determina si el periódo de publicación de la noticia correspende con el filtro establecida y prepara el contenido para su clasificación.
- Clasificación en secciones. Determina a que sección pertenece la noticia basado en su contenido.
- Visualización de enlaces a las noticias. En este módulo se muestran vínculos a las noticias clasificadas que cumplen con los filtros de periodo y secciones establecidos por el usuario.

Al finalizar este trabajo se contará con:

- Documentación de análisis y diseño.
- Recolector de noticias.
- Manual técnico y de usuario.

5. Metodología

La metodología que se utilizará para la realización del presente trabajo terminal será la incremental [9], debido a que uno de los objetivos es el crecimiento progresivo, es decir se realizan entregas parciales en un periodo de tiempo corto y así reducir riesgos en el proyecto, como se puede ver en la Figura 2.

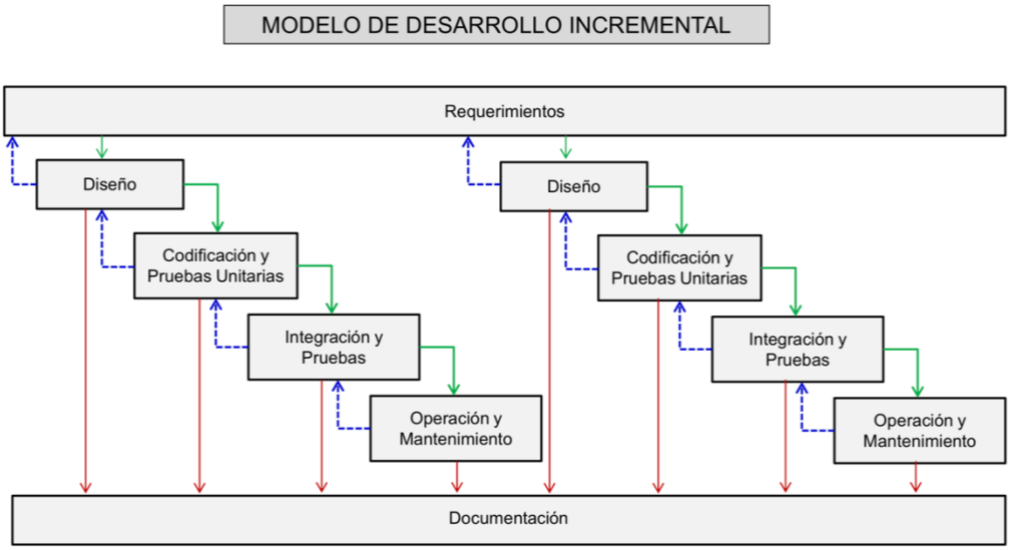

Figura 2. Modelo de desarrollo incremental [7].

6. Cronograma

Nombre del alumno(a): Hernández Gómez Carlos Andrés
Título del TT:

Actividad	FEB	MAR	ABR	MAYO	JUNI	AGO	SEP	OCT	NOV	DIC
Selección de fuentes de información										
Definición de requerimientos										
Análisis y diseño del módulo recolector de noticias										
Implementación del módulo recolector de noticias										
Pruebas del módulo recolector de noticias										
Evaluación de TT1										
Análisis y diseño del extractor de información										
Implementación del extractor de información										
Integración de los módulos										
Pruebas integrales										
Redacción de los manuales de usuario y técnico										
Evaluación de TT2										

Nombre del alumno(a): Meza Martínez Luis Daniel
Título del TT:

Actividad	FEB	MAR	ABR	MAYO	JUNI	AGO	SEP	OCT	NOV	DIC
Definición de secciones a utilizar en el filtrado.										
Definición de requerimientos.										
Adaptación del módulo clasificador de noticias.										
Pruebas del módulo clasificador de noticias.										
Evaluación de TT1										
Análisis y diseño del módulo visualizador de noticias filtradas.										
Implementación del módulo visualizador de noticias filtradas										
Integración de módulos.										
Pruebas integrales.										

Redacción de los manuales de usuario y técnico										
Evaluación de TT2										

7. Referencias

[1] S/A. (S/A). Importancia de las Noticias. 07/09/2018, de INNOVACION INTERNAUTICA Sitio web: https://innovainternetmx.com/2014/12/importancia-de-las-noticias/

[2] Manning, C., Raghavan, P. and Schütze, H. (2009). Introduction to information retrieval. New York: Cambridge University Press, pp.443-459.

[3] Bernabeu Morón, N. (2013). La noticia y el reportaje. España: Ministerio de Educación de España, p.9.

[4] MIKE MCCUE. (S/A). FLIPBOARD ES TU MOMENTO. 06/09/2018, de FLIPBOARD Sitio web: https://es-es.about.flipboard.com/?noredirect=es_ES

[5] HuffPost. (S/A). HuffPost México. 06/09/2018, de HuffPost Sitio web: https://www.huffingtonpost.com.mx/p/huffpost-mexico-about-us

[6] Google. (S/A). Google News. 12/09/2018, de Google Sitio web: https://news.google.com/

[7] J. García ,L. Ramírez, M. Sanchéz, "Clasificación de noticias de diarios de circulación nacional mediante aprendizaje automático", trabajo terminal, ESCOM IPN, 2018.

[8] El Economista Y ComScore. (03/09/2018). Ranking de Medios Nativos Digitales. 12/09/2018, de El Economista Sitio web: https://www.eleconomista.com.mx/tecnologia/Ranking-de-Medios-Nativos-Digitales-20170830-0177.html

[9] D. Tapias, "Proyectos de desarrollo software", 2014, [En linea]. Disponible en : http://arantxa.ii.uam.es/~proyectos/teoria/C5_Proyectos%20de%20desarrollo%20software.pdf

8. Alumnos y Directores

CARÁCTER: Confidencial
FUNDAMENTO LEGAL: Art. 3, fracc. II, Art. 18, fracc. II y Art. 21, lineamiento 32, fracc. XVII de la L.F.T.A.I.P.G.
PARTES CONFIDENCIALES: No. de boleta y Teléfono.

Hernandez Gomez Carlos Andres.- Alumno de la carrera de ingeniería en sistemas computacionales en la Escuela Superior de Cómputo del Instituto Politécnico Nacional, Boleta: 2015620193, Tel. 5546218045, email: carlosandreshg.ipn@gmail.com

Firma:_____

Meza Martínez Luis Daniel.- Alumno de la carrera de ingeniería en Sistemas Computacionales en la Escuela Superior de Cómputo del Instituto politécnico nacional , Boleta:2015630305 , Tel.5573994181 , email: *ldanielmezam@gmail.com*

Firma:_____

Joel Omar Juárez Gambino.- Licenciado en Informática por la Facultad de Informática, UAS. Maestro en Ciencias de la computación por el CIC, IPN. Sus áreas de estudios son: Inteligencia Artificial, Lenguaje Natural y Representación de conocimiento. Departamento de Ciencias e Ingeniería de la computación, ESCOM, Tel. 57296000 Ext. 52022, email: omarjg@gmail.com

Firma:_____

Consuelo Varinia García Mendoza.- Ingeniera en Sistemas Computacionales por la ESCOM, IPN , UAS. Maestra en ciencias en Tecnología Avanzada por el CICATA-Legaria, IPN, Doctora en Técnologias Avanzadas por la CICATA-Legaria, IPN. Sus áreas de estudios son: Análisis de algoritmos y Optimización. Departamento de Ciencias e ingeniería de la computación, ESCOM, Tel. 57296000 Ext. 52022, email: consuelo.varinia@gmail.com

Firma:_____