Generación automática de contenido

Authomatic generation content

Juan David Osorio Ortiz

*Universidad Tecnológica de Pereira, Pereira, Colombia*

Juandavid.osorio1@utp.edu.co

***Resumen*— En este paper se plantea y se muestra como es el proceso para trabajar con una IA generadora automática de contenido, en el cual se explica cómo es el paso a paso desde el momento en que se está haciendo “web scraping” para seleccionar los datos, hasta el momento en que se tenga la estructura completa.**

***Palabras clave—* Generación automática de contenido, automática, información, web scraping, bases de datos, scraping.**

***Abstract*— In this paper, the process to work with an automatic content generating AI is presented and shown, which explains how it is the step by step from the moment “web scraping” is being done to select the data, until the moment you have the complete structure.**

***Key Word ---* Automatic content generation, automatic information, web scraping, databases, scraping.**

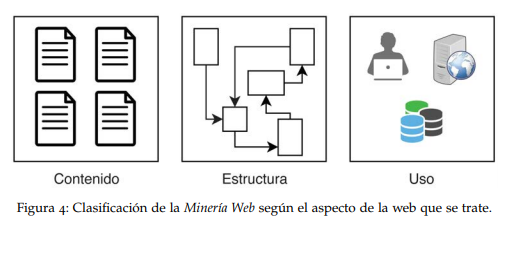
1. INTRODUCCIÓN

En la actualidad y durante las últimas décadas el internet a facilitado la generación y el acceso a exorbitantes cantidades de información, no solo de nuestros alrededores si no información de cualquier tema de todo el planeta en general. Por lo que si tuviéramos que buscar algo sobre un tema en específico nos tocaría leer miles y miles de publicaciones para encontrar lo que nosotros exactamente necesitamos. Esta labor en si nos tomara horas y horas de investigación para poder encontrar la información adecuada.

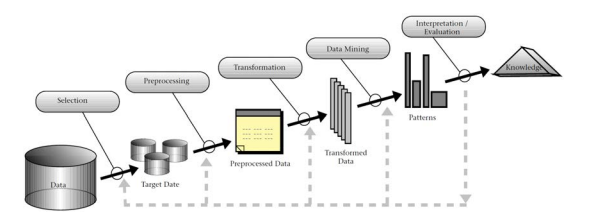
A partir de esta problemática de generación de contenido sin restricción y la necesidad del ser humano de estar investigando sobre temas específicos que le interesan se plantea la generación automática de contenido una rama de la inteligencia artificial encargada de filtrar información según las necesidades de las personas ya que el 90% de la información que encontraríamos en una búsqueda normal sería información inservible para nosotros, o que no es de nuestro interés. La labor de la IA en este caso es mostrarnos el 10% esencial de la información que estamos buscando.

1. CONTENIDO

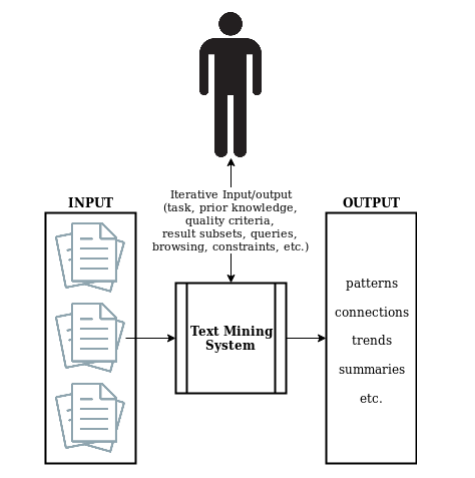
Partiendo del problema que se plantea para generar contenido automáticamente se tienen que seguir una serie de tapas para solucionarlo.



La etapa inicial que hay que tener en cuenta para la generación de contenido automático es realizar un proceso arduo de minería y extracción de datos de la WWW y de repositorios de información encontrados en esta. Para esto se usan diferentes metodologías como el text mining basado en extraer información solo de texto, hay otras técnicas mas generales como el web mining, tiene muchas similitudes con el text mining pero a diferencia del anterior mencionado este no excluye ningún tipo de datos, desde documentos de texto, imágenes, videos, archivos de audio, hipervínculos y cualquier otra clasificación de la información que se tenga.



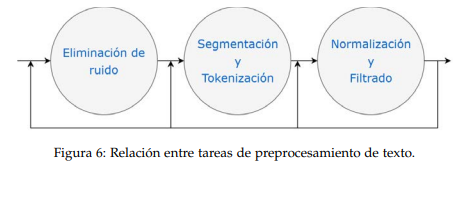
Una vez se tiene un gran repositorio de información todo esto se manda a un sistema de minería de datos, estos reciben toda la información y durante el proceso nos da como salida patrones, conexiones, tendencias y resúmenes entre otros, esto se puede explicar mejor con el siguiente esquema.



Debido a que toda la información que se obtiene en la etapa de minera esta en formato web y tiene muchas capas de html y css por encima seguir 3 pasos para poder completar la recolección y selección de la información:

* Recuperar datos de html a través de una Uniform Resource Locator (URL)
* Analizar los datos para almacenarlos.
* Opcionalmente se puede continuar con otra pagina para repetir este procedimiento.

Una vez se cumple con este primer paso, se continua con el segundo paso, este es la representación del texto obtenido en la etapa anterior, esto para que pueda ser tratado por los algoritmos posteriores. Para esto se tienen que seguir un conjunto de tareas como la eliminación del ruido, segmentación, tokenizacion, normalización y reducción del léxico del texto.



* Eliminación del ruido: Teniendo en cuenta que el contenido recopilado es html, la eliminación del ruido consiste en la eliminación de encabezados, pies de página, etiquetas css y html o cualquier otro tipo de metadato.
* Segmentación: Es la parte que divide todo el texto limpio de un documento en capítulos, párrafos, secciones, oraciones, palabras y en algunos casos silabas.
* Normalización: Este proceso radica en evitar las diferentes formas de entender cómo se expresa una palabra, por sus diferentes conjugaciones y usos en distintas oraciones.
* Filtrado: Se tiene como objetivo filtrar todos los tokens que no cumplan con ciertas condiciones

Al finalizar este proceso se sigue con la generación de resúmenes automáticos de todos los documentos, se pueden generar resúmenes extractivos y por abstracción.

Una vez finalizada esta parte se sigue con la representación del documento, esta representación depende de los problemas a resolver, No es lo mismo querer extraer palabras clave, clasificar documentos, resumirlos o incluso analizar sentimientos.

Una vez se tienen todos estos pasos ya tenemos la mayoría de los problemas resueltos lo último seria organizar todo en una plataforma en la que se pueda acceder a toda la información procesada y generada automáticamente.

1. CONCLUSIONES

* La generación de contenido automática es uno de los mejores avances en las ultimas décadas debido a que en la web se están generando miles de artículos a diario, y poder acceder a la información que necesitamos sin necesidad de leer antes una gran cantidad de información innecesaria es un punto a favor de esto.
* Cuando se habla de generación automática de contenido se esta hablando de muchas áreas de la ciencia y la computación en la que se aplica con esta metodología.

REFERENCIAS

* 1) Generación automática inteligente de resúmenes de textos con técnicas de Soft Computing, Augusto Villa Monte.