MODELO DE ANÁLISIS DE INFORMACIÓN DESESTRUCTURADA UTILIZANDO TÉCNICAS DE RECOPILACIÓN Y MINERÍA WEB

DIRECTOR: Karanik, M. J. **INVESTIGADORES:** Suénaga, R. Favret, F. Eckert, K. B. **COLABORADORES:** Rojas, M. Pfeifer, H.

RESUMEN

Cuando se hacen búsquedas de información en internet, saber exactamente lo que se busca es primordial para localizar rápidamente lo que se requiere. Pero cuando lo que se necesita es identificar qué documentos hay acerca de un tema en particular (sin identificar su denominación precisa), los buscadores identifican infinidad de sitios y documentos que hace casi imposible revisarlos individualmente hasta dar con aquellos que se acercan a lo que al usuario le interesa.

Esa situación requiere de soluciones especializadas que utilicen técnicas de interpretación de contenido de páginas web y documentos de internet que, a partir de una especificación de lo requerido por el usuario, busque, analice y clasifique la información disponible de acuerdo a lo solicitado.

Este trabajo aborda la situación descripta precedentemente, a partir del cual se propone desarrollar y utilizar un modelo para el proceso de identificación de sitios y documentos, basado en la integración de técnicas de recopilación, exploración y análisis de información en la web.

Primeramente se describen las características de los procesos de identificación de información en internet, como así también las métricas de evaluación que permiten identificar la relevancia de la información de interés.

En el desarrollo del trabajo se propone un modelo correspondiente a un proceso que consiste en la utilización de resultados de los buscadores de internet (Google, Bing, MSXML Excite e Intelligo), a partir del cual se desencadena un proceso de exploración de los enlaces de cada sitio identificado, para luego proceder a asignar puntajes de acercamiento a los requerimientos del usuario (uno de los parámetros significativos es asignado por el análisis semántico), finalmente se ordenan los documentos de acuerdo a los valores asignados (raniking).

El modelo se probó en dos escenarios, uno referido a información sobre herramientas de educación digital y el segundo referido a información sobre técnicas vinculadas con seguridad informática. Las pruebas del modelo proporcionaron ordenamientos que ubicaron mejor a los recursos más relevantes en los rankings construidos, los que fueron validados por usuarios especializados que configuraron los escenarios de búsqueda inicial.

Palabras clave: Minería web; análisis semántico; recuperación de información.

INTRODUCCIÓN

Durante mucho tiempo las dificultades de acceso a las fuentes de información fue el factor preponderante para conseguir datos fiables, pero con el avance de las tecnologías de la información y las comunicaciones (TICs) este inconveniente fue desapareciendo dando lugar a otro con el mismo efecto: la sobresaturación de información. Por ello, existen grandes volúmenes de información que no pueden ser manejados con métodos tradicionales debido al nivel de desestructuración, su disponibilidad en distintos formatos, que se encuentran parcial o totalmente desconectados y están altamente distribuidos.

El área de análisis de grandes volúmenes de información ha tomado una relevancia excepcional dentro de las TICs. Esto se debe a que la tecnología, a medida que evoluciona, facilita el desarrollo de algoritmos para tareas de análisis de datos, tales como clasificación, búsqueda de patrones, determinación de tendencias, construcción de modelos descriptivos y predictivos, entre otras. Como consecuencia, las técnicas y algoritmos son cada vez más eficientes y producen resultados más precisos, convirtiéndolos en herramientas apropiadas para la obtención de información útil.

En base a lo expuesto antes, este proyecto abarca el estudio e implementación de técnicas de búsqueda y análisis de información útil. Específicamente se busca modelar un sistema integral de información que incluya investigación sobre sistemas de recopilación de necesidades, búsqueda automática, exploración y minería web y herramientas de toma de decisiones.

RECUPERACIÓN DE INFORMACIÓN Y RECUPERACIÓN DE DATOS

Existen diferencias marcadas entre la recuperación de información (RI) y la recuperación de datos (RD). Estas diferencias están relacionadas a los tipos de objetos con los que trata cada una, la representación de estos objetos, la especificación de las consultas y los resultados que se obtienen. En primer lugar, la RD trata con valores y claves de búsqueda con una estructura definida, mientras que la RI debe lidiar, en ambos casos, con las dificultades del procesamiento del lenguaje natural. Por otro lado, se puede ver a la RD como una aproximación a la RI, en las situaciones en que se busca determinar los recursos de una colección que contienen las palabras de la clave de búsqueda ingresada por el usuario. Sin embargo, desde el punto de vista de la RI, lo más probable es que los resultados obtenidos sean irrelevantes para la necesidad de información que el usuario posee, ya que la ocurrencia de palabras de la

clave de búsqueda en un recurso, no es suficiente como para afirmar si éste es relevante o no (Baeza-Yates y Ribeiro-Neto, 1999).

RELEVANCIA Y MÉTRICAS DE EVALUACIÓN

Determinar si un recurso es relevante es una actividad compleja, debido a que se trata de un juicio subjetivo, lo que significa, que dos usuarios pueden determinar niveles de relevancia distintos para un mismo recurso. Por lo tanto, se hace evidente que la relevancia consiste en realidad en la consideración de varios factores, entre los que se encuentran las características del recurso, las características de la necesidad de información y la subjetividad del usuario que elabora la clave de búsqueda (Tolosa et al. 2008).

Con respecto a esto, en Blair D. (1990) se afirma que es más fácil llevar a cabo la determinación de la relevancia de un recurso determinado, que explicar cómo fue obtenida o qué criterios se utilizaron.

Con el objetivo de evaluar, que tan acertada es la respuesta proporcionada al usuario y, por lo tanto, que tan efectiva es la determinación de relevancia del sistema de recuperación de información (SRI), es necesario utilizar un conjunto de métricas que permitan llevar a cabo dicha labor. Existen varias métricas que permiten evaluar la calidad de los resultados obtenidos, de las cuales las más utilizadas son la precisión y exhaustividad, planteadas en Cleverdon C. (1966).

La precisión se define como la proporción de los recursos recuperados que son relevantes y permite evaluar la habilidad del sistema para generar un ranking en el que las primeras posiciones estén ocupadas por recursos relevantes (Landauer y Laham. 1998).

La exhaustividad se define como la proporción de los recursos relevantes que han sido recuperados y permite evaluar la habilidad del sistema para encontrar todos los recursos relevantes de una colección (Baeza-Yates y Ribeiro-Neto, 1999).

MODELOS PARA RECUPERACIÓN DE INFORMACIÓN

Existen varios enfoques de modelos de RI, que representan distintos puntos de vista. Gran parte de estos, se centran en el cálculo de la probabilidad de que el contenido de los recursos se relacione al contenido de la clave de búsqueda. Algunos modelos de RI más representativos, son: el modelo Booleano, el Vectorial, el Probabilístico y el enfoque Semántico.

MODELO PROPUESTO

En Eckert et al. (2016) y Favret et al. (2016) se describe un modelo de RI que tiene dos objetivos bien definidos: el primero consiste en capturar de manera precisa la necesidad de información que posee el usuario y presentar los resultados obtenidos. El segundo, consiste en la implementación de métodos y técnicas que permitan la identificación, recuperación y análisis continuo de recursos existentes en la Web. En principio, los dos objetivos mencionados, son implementados mediante el *Módulo de Recopilación de Requerimientos* (MRR) y el *Módulo de Minería Web* (MMW), respectivamente.

Con el fin de implementar un sistema que lleve a cabo la determinación de la relevancia de recursos, mediante Técnicas Semánticas (TS) ya implementas en [6] [7] [8], se define la arquitectura del sistema, que se presenta en la Figura 1.

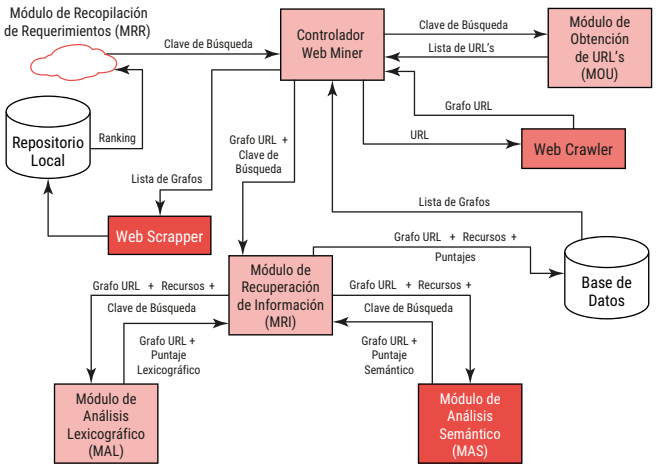

Figura 1. *Arquitectura léxico-semántica*

El proceso de análisis de recursos y generación de rankings comienza cuando el MRR envía la clave de búsqueda ingresada por el usuario al Controlador Web Miner, que se encarga de coordinar el funcionamiento de todo el sistema. Esta clave, se pasa al Módulo de Obtención de URL's (MOU), donde se genera una lista única de URL's conformada por los primeros diez resultados de los siguientes buscadores: Google, Bing, MSXML Excite e Intelligo. Confeccionada dicha lista, se la retorna al Controlador Web Miner, donde se descartan las repetidas e inicia el proceso de análisis.

Cada una de las URL resultantes, son representadas mediante nodos, con estado de procesamiento "No Explorado". La representación mediante nodos permite mantener una estructura de datos por cada URL, compuesta por los puntajes procedentes de los análisis, el contenido de dicha URL (recurso Web) y el estado de procesamiento, que puede ser "Explorado" o "No Explorado".

La siguiente operación consiste en enviar uno de los nodos marcado como "No Explorado" al módulo Web Crawler, donde se descubren los enlaces directamente relacionados a su URL y se construye un grafo acíclico dirigido que represente estas relaciones. En este grafo, el nodo recibido es la raíz y los enlaces descubiertos se representan como sus nodos hijo. Como resultado de esto, se cambia el estado de procesamiento de la raíz a "Explorado" y se retorna el grafo completo al Controlador Web Miner.

Seguidamente, se envía el grafo generado y la clave de búsqueda, al Módulo de Recuperación de Información (MRI), donde por cada nodo se obtiene el contenido del mismo (HTML o PDF) y se ejecuta el análisis de relevancia, que determina su grado de correlación con respecto a la clave de búsqueda. Para el análisis de relevancia se utilizan dos técnicas: las TL implementadas mediante el Módulo de Análisis Lexicográfico (MAL) y las TS implementadas mediante el Módulo de Análisis Semántico (MAS).

Como resultado de la ejecución de estos módulos, por nodo, se obtienen tres puntajes correspondientes a las técnicas CRank, Okapi y VSM, y un cuarto puntaje que representa a la relevancia obtenida mediante las técnicas semánticas.

Al finalizar estos análisis, se almacena en la base de datos los nodos del grafo analizado, es decir, los puntajes, el contenido de la URL y el estado de procesamiento de cada nodo.

Posteriormente, en el Controlador Web Miner, se comienza con la presentación de los resultados. Para ello, inicialmente, se genera una lista compuesta por los grafos obtenidos hasta el momento.

Esta lista se envía al módulo Web Scrapper, donde se calcula el puntaje final de cada nodo mediante el Modelo de Integración Léxico – Semántico (MILS) y se confecciona el ranking de recursos a ser presentado al usuario.

Este ranking consta de cincuenta posiciones, con el fin de limitar la cantidad de información presentada al usuario. Para generarlo, en principio, se agrupan los recursos de acuerdo su dominio.

Ya con los grupos conformados, se determinan las cincuenta posiciones del ranking, considerando los puntajes arrojados por el MILS de los recursos principales de cada dominio.

Por cada posición del ranking, se genera un archivo JSON que contiene el puntaje de relevancia correspondiente al recurso principal, la posición que ocupa en el ranking y las URL's de los recursos relacionados al mismo. Dichos archivos son almacenados en un repositorio local, permitiendo que sean recuperados por el MRR, para presentarlos al usuario.

El proceso continúa con la expansión de los demás nodos marcados como "No Explorado". Finalizada la primera iteración, se procede a verificar si existen nodos obtenidos a partir del descubrimiento de enlaces relacionados. De ser así, se da inicio a una nueva iteración, lo que implica el descubrimiento de nuevos nodos y la realización del análisis de relevancia por cada uno de ellos, resultando en una reestructuración del ranking a ser presentado al usuario. En caso contrario, se da por finalizado el proceso de análisis de recursos y generación de rankings. Además, cabe destacar que el usuario puede finalizar dicho proceso cuando desee, lo que supone otro punto de corte para la ejecución del sistema.

MÓDULO DE ANÁLISIS SEMÁNTICO

El objetivo de este módulo es implementar métricas de relación y similitud semántica, que contribuyan a determinar la relevancia de los recursos analizados. El esquema general de este módulo se presenta en la Figura 2.

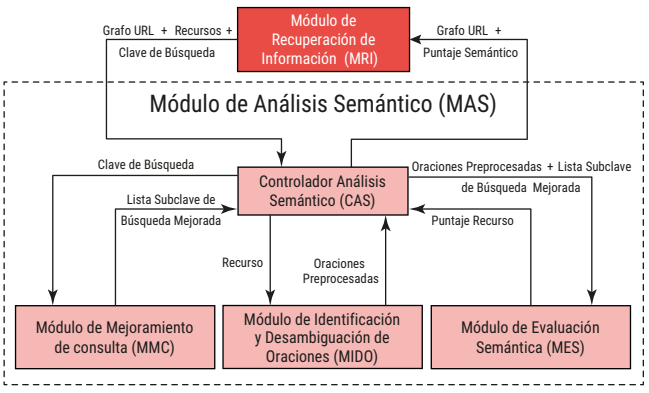

Figura 2. *Esquema del MAS*

Como se puede apreciar, el MAS, está compuesto a su vez por un conjunto de módulos que interactúan entre sí, donde el Controlador Análisis Semántico (CAS) es el encargado de coordinar toda su operatoria.

El proceso de análisis semántico comienza al recibir del MRI el Grafo URL a analizar y la clave de búsqueda, donde cada nodo del grafo está compuesto por su correspondiente recurso.

En primera instancia, el CAS envía la clave de búsqueda al Módulo de Mejoramiento de Consulta (MMC), donde se eliminan errores ortográficos, los stopwords[1] y se identifica el sentido de los términos que la componen. Además, se la segmenta en subclaves, lo que permite que se tenga en cuenta la importancia de las distintas partes de la clave de búsqueda. Como resultado, se retorna la lista de subclave de búsqueda mejorada al CAS.

Luego se envía un recurso correspondiente a un nodo, al Módulo de Identificación y Desambiguación de Oraciones (MIDO), donde se segmenta a su contenido en las oraciones que lo conforman. Por cada oración, se desambigua el sentido de las palabras que la componen. Como resultado, se obtiene una lista de oraciones preprocesadas, que se retorna al CAS.

A continuación, se envía esta lista de oraciones preprocesadas junto a la lista subclave de búsqueda mejorada al Módulo de Evaluación Semántica (MES), donde se aplica la métrica de relación y similitud semántica para determinar el puntaje de relevancia correspondiente al recurso analizado. Finalmente se retorna el puntaje recurso al CAS.

Hecho esto, se comprueba si existen nodos por analizar en el grafo, de ser así, se vuelve a realizar el mismo procedimiento. En caso contrario, se finaliza el análisis semántico, retornando al MRI el Grafo URL junto a los puntajes semánticos correspondientes.

CONCLUSIONES

Para realizar las pruebas y analizar los resultados del modelo, se consideraron dos escenarios. El primero corresponde al ámbito de la educación digital, mediante la utilización de la técnica "Digital Storytelling". El segundo escenario correspondió al área de la seguridad informática, más precisamente en relación a técnicas de ataques por envenenamiento de cookie (Cookie Poisoning).

Durante el desarrollo del trabajo se persiguió el objetivo de analizar las técnicas propuestas, lo que permitió evaluar distintos aspectos relacionados a la recuperación, la generación de rankings, y los criterios de las técnicas semánticas. Las pruebas de recuperación muestran la efectividad de las técnicas semánticas, obteniendo recursos relevantes desde internet, descartando los no relevantes.

Se logró verificar la efectividad en la determinación de la relevancia de documentos específicos sobre el conjunto infinito de recursos existentes en la web, lo cual incide en la cantidad de recursos relevantes presentados al usuario.

Las pruebas de ordenamiento, por otro lado, se centraron en la capacidad de cada técnica de ubicar mejor a los recursos relevantes en los rankings que construyen. En una situación ideal, los recursos más relevantes debieran ubicarse en las primeras posiciones del ranking. Sin embargo, esto no siempre se presentó así en la práctica, debido a la complejidad inherente a la determinación de la relevancia.

Se llevaron a cabo las pruebas sobre el MILS, que permitieron observar los resultados del ordenamiento de recursos a partir de la combinación

1 Stopword: Palabras vacías, o comúnmente utilizadas y que no contribuyen al significado.

de los criterios considerados por las técnicas semánticas y comprobar como la consideración de distintos enfoques, puede contribuir a la correcta estimación de la relevancia.

En definitiva, las pruebas realizadas proveyeron de elementos de análisis considerando distintas perspectivas que permitieron observar y destacar distintos aspectos relacionados al comportamiento del modelo.

BIBLIOGRAFÍA

• Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern information retrieval*. New York : Harlow, England: ACM Press ; Addison-Wesley.

• Blair, D. C. (1990). *Language and representation in information retrieval. Amsterdam;* New York: New York, N.Y., U.S.A: Elsevier Science Publishers; Distributors for the U.S. and Canada, Elsevier Science Pub. Co.

• Cleverdon, C., Mills, J. and Keen, M. (1966) ASLIB Cranfield Research Project: factors determining the performance of indexing systems.

• Eckert, K., Alvarenga, V. M., Barboza, M., Witzke, L. M., and Araldi, L. (2016). *Vigilancia tecnológica e inteligencia competitiva basada en técnicas de minería de la web*, presented at the XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016).

• Eckert, K., Favret, F., Barboza, M., Witzke, L. M. and Alvarenga, V. M. (2016). *Modelos de análisis de información para la toma de decisiones estratégicas del sector tealero*, presented at the XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina)

• Favret, F., Montiel, R., Alvarenga, V., Barboza, M., and Witzke L. (2016). *Recuperación de información basada en técnicas de minería Web*, Pag. 7.

• Landauer, T. K., Foltz, P. W. and Laham, D. (1998). *An introduction to latent semantic analysis*, Discourse Process., vol. 25, no. 2–3. Pag. 259–284.

• Tolosa, G. H. and Bordignon, F. R. A. (2008). *Introducción a la Recuperación de Información*. Tolosa y Bordignon.